Automatic Personality Recognition from Non-verbal Acoustic Cues: Bridging the Gap Between Psychology and Computer Science

Dina Abdulaziz Al-Hammadi

Supervisor: Prof. Roger K. Moore

Department of Computer Science

The University of Sheffield

This thesis is submitted for the degree of
Doctor of Philosophy

Speech and Hearing Research Group

September 2021

I would like to dedicate this thesis to my loving parents, my beloved husband, and my darling children for their love and support...

Declaration

I hereby declare that I am the sole author of this thesis, except where specific reference is made to the work of others. The contents of this thesis are my original work and have not been submitted for any other degree or any other university. Some parts of the work presented in Chapter 6 have been published in conference proceedings as follows:

- D. Al-Hammadi and R. K. Moore, "*Using Sampling Techniques and Machine Learning Algorithms to Improve Big Five Personality Traits Recognition from Non-verbal Cues*", 2021 National Computing Colleges Conference (NCCC), 2021, pp. 1-6.

Dina Abdulaziz Al-Hammadi
September 2021

Abstract

Human-computer interaction (HCI) is an evolving research field; it has changed from focusing on usability and interface design to more complex interaction and adaptivity design. Many disciplines have become involved in HCI including psychology. One interesting aspect of psychology, which is relevant to HCI is personality. Personality is a stable pattern of behaviour and thought that uniquely characterizes individuals and how they behave in social contexts. There are many personality theories but the big five, introduced in 1980 by Lewis Goldberg, is the most successful and widely used. The big five are openness, conscientiousness, extraversion, agreeableness, and neuroticism, known as OCEAN. In the last ten years, there has been a growing interest in 'social signals' since its introduction by Alex Pentland in 2007. Social signals were later defined in 2010 by Poggi and D'Errico as "*communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions, social emotions, social attitudes and social relationships*". Social signals are non-verbal cues, such as face features, body gestures, and vocal behaviour. Other researchers have shown that social signals can be successful at predicting the behavioural outcomes of social situations such as speed dating. Hence, since personality affects behaviour, social signals should be correlated with personality. There are proposed methods to identify the big five personality traits for personality recognition, especially acoustic cues from speech. However, such methods are focused mainly on personality recognition by strangers (zero-acquaintance) and not accurate personality recognition. The research shows the difference between personality perception and personality recognition and demonstrate how to recognise personality accurately. Further research into personality recognition has unveiled a huge gap in personality research between computer science community and psychology community. Available corpora are built on stranger agreement (personality perception), and not on accurate personality judgement (personality recognition). Therefore, new corpus was collected based on accurate personality judgement model and experiments with the new corpus show that there is a correlation between the big five and social signals (acoustic cues). In addition, the research shows that social signals (acoustic cues) can be used to recognize all big five personality traits as opposed to perceiving them. The research has found that by providing valid and accurate personality data then social signals can be captured and it is possible to have accurate automatic personality recognition. The results demonstrate how social signals are identified, captured and analysed to recognize personality traits accurately from speech alone. This research anticipates its results to be of value to many HCI research areas such as healthcare and e-learning. Furthermore, personality recognition is a major issue in human-robot interaction (HRI), and this research will be of relevance to their future developments.

Acknowledgment

First and foremost, I would like to thank Allah for giving me the strength, motivation and perseverance to finish this amazing journey. I am proud of myself for accomplishing this thesis. I am forever blessed to have my family around me and always supporting me to keep going and never give up. To my mother Hind and my father Abdulaziz, thank you for your unconditional love and support. You taught to to always dream and aim high and achieve the best because I can do it. You provided me with everything that I needed and I cannot ask for more. Thank you for giving me the courage to keep going, the love for education and research and be the best version of myself. I am forever grateful and I am lucky to be your first born. To my sisters and brothers: Faisal, Lina, Dima , Rima, Yara, Lama and Mohammad; thank you for always being by my side and surrounding me with much love and support especially when I was feeling down. I'm blessed to have you around me. To my love, my husband Abdulrahman and my moon and stars, Abdulaziz, Omar, Fahad and Aljoharah. This journey was fun because I had you around me. Through thick and thin, your smiles and laughters made me stronger. Thank you for your patience. During this journey I have made great friends who stood by me and knew what it was like to face the hardships; my wonderful support system: Aisha AlArfaj, Fatimah AlHayyan and Maha AlSweilem.

Special recognition to my supervisor professor Roger K. Moore, I remember our first meeting and my goal was to make a contribution that makes you proud of me as a researcher, and in our last meeting you said that you were proud. I am happy to have achieved such recognition from a professor of your calibre. I am forever grateful for all the lessons you have taught me. I know I'm on the way to become a greater academic and supervisor because of you. To my wonderful panel Dr. Mark Stevenson and Prof. Kalina Bontcheva, although our meetings were few, I'm always intrigued by the input you give me and the many possibilities I can explore. Thank you for your great feedback.

This journey would not have started if it were not for my internal supervisor Dr. Hmood Al-Dossari. The article you have sent me was the catalyst for this research and this thesis. Thank you.

My hero in the psychology community, the esteemed Prof. Tara Letzring whom I'm in debt for the knowledge and time you have given me to discuss in depth the science behind personality psychology.

Last but not least, I would like to thank my country Saudi Arabia and my University; Princess Noura Bint Abdulrahman University. Thank you for the Deanship of Graduate Studies and Scientific Research's continuous support: Dr. Fatima Al-Shehri, Dr. Mashael Al-Mufirej, Dr. Haya Al-Otaibi, and Dr. Hadil Shaiba. Thank you PNU for the scholarship to pursue my dream of higher education at the respected University of Sheffield which I'm proud to be an alum.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The beginning is the most
important part of any work

Plato

Human-computer-interaction (HCI) is an evolving and expanding research field (Carroll, 1997). When computers were first introduced, only computer experts were able to interact with them (Booth, 2014; Shneiderman, 1981; Helander, Landauer, & Prabhu, 1997). Computers were used to facilitate and improve the work process in companies and organizations. Therefore, the focus was on a user-friendly interface (Helander et al., 1997; Preece, Sharp, & Rogers, 2015). This has led to dedicated practices, techniques, and methodologies for user-interface design (Booth, 2014; Carroll, 1997). As technology has evolved and new technologies have been introduced, the scope has expanded from only a user-friendly interface to include user experiences, such as joy, happiness, and motivation in addition to efficiency and productivity (Lazar, Feng, & Hochheiser, 2010; MacKenzie, 2012; Rogers, 2004). Human-centred design (HCD) was proposed to address the technological challenges (D. Norman, 2013). The research continues to grow, and "*the study of HCI is now effectively a boundless domain*" (Barnard, May, Duke, & Duce, 2000).

In the 1970s, HCI involved a combination of software engineering and human factors (R. Mayer, 1975). However, human factors were considered at the final stage of the development process, which led to minor ineffective changes in the software (Carroll, 2003). A few years later, human factors (ergonomics) have included the cognitive sciences (Shneiderman, 1979, 1983). Since then, many disciplines have become involved with HCI, such as artificial intelligence, software engineering, and psychology (Preece et al., 2015; Rogers, 2012).

The human factor has represented a major design dilemma for researchers (Dix, 2004). Human attributes, such as race, language, sex, attitude, emotion, personality, and age, are what makes humans complicated (Shneiderman & Plaisant, 2010). An important aspect of psychology that is relevant to HCI is personality (Pianesi, 2013). When a system is able to understand a human's individual characteristics and behaviour, it can adapt to their needs and capabilities, leading to better efficiency and higher productivity (Nass & Brave, 2007; D. Norman, 2013; Siegert, Haase, Prylipko, & Wendemuth, 2014).

Allport (1961) defined personality as "*...the dynamic organization within the individual of those psychophysical systems that determine his characteristics behaviour and thought*". There are many personality theories, but the big five introduced by Lewis Goldberg in 1980 (Goldberg, 1980, 1990) is the most successful and widely used theory of personality recognition (Celli, 2012; McCrae & Costa, 1987). The big five personality dimensions are openness, conscientiousness, extraversion, agreeableness, and neuroticism (Carducci, 2009; John, Robins, & Pervin, 2008), collectively known as OCEAN.

There are many proposed methods to identify the big five personality traits (Alam & Riccardi, 2014b; Farnadi, Zoghbi, Moens, & De Cock, 2013; Farnadi et al., 2014; Nowson & Gill, 2014; Verhoeven, Daelemans, & De Smedt, 2013). However, those methods focus mainly on personality perception from strangers (zero-acquaintance) (Lepri, Kalimeri, & Pianesi, 2010; Ivanov, Riccardi, Sporka, & Franc, 2011; Batrinca, Lepri, Mana, & Pianesi, 2012a; Valente, Kim, & Motlicek, 2012a; Celli, Bruni, & Lepri, 2014; Nowson & Gill, 2014). In addition, existing corpora are not appropriately designed for personality recognition but rather for personality perception from strangers. Therefore, the ground truth associated with the corpora are not suitable for personality recognition. This study aims to design and build a corpus based on an accurate personality model of ground truth—a personality recognition corpus. It aims to answer the following question: can social signals be used for big five personality trait recognition?

In the last ten years, there has been a growing interest in 'social signals', introduced by Pentland (2007b). Social signals are "*communicative or informative signal or a cue which, either directly or indirectly, provides information about 'social facts' that is, about social interactions, social emotions, social attitudes, evaluations and stances, social relations, and social identities*" (Poggi & D'Errico, 2012). Several experiments have shown that social signals can be successful at predicting the behavioural outcomes of social situations, such as speed dating (Pentland & Heibeck, 2010). Personality affects behaviour and how humans react in social contexts (Cassell, 2000a). Therefore, social signals are used to recognise personality by capturing behaviour.

Interestingly, interest in personality traits has been recently growing in several various fields. In psychology research, psychologists have become interested in studying human behaviour and how it can be used to predict future behaviour in adulthood life scenarios (Taggart, Bannon, & Hammett, 2019; Stoll et al., 2020; Bühler, Finkenauer, & Grob, 2020) and predict future life goals (Reisz, Boudreaux, & Ozer, 2013).

Businesses are investing in user-related research. Understanding user behaviour and having the ability to predict how it affects consumer products and services (Whelan & Davies, 2006). An article published by Xu, Frey, Fleisch, and Ilic (2016) explores the relation between personality traits and consumer mobile apps. Mønsted, Mollgaard, and Mathiesen (2018) study the relation between personality traits and phone usage.

In business research, some studies explored the link between entrepreneurial traits and personality traits (Leutner, Ahmetoglu, Akhtar, & Chamorro-Premuzic, 2014). Other research studied personality traits link to the labour market and how it matched employees with prospective employers (Haylock & Kampkötter, 2019) and employee satisfaction and attitude toward assigned tasks (Rubenstein, Zhang, Ma, Morrison, & Jorgensen, 2019). Other studies focused on human behaviour prediction for commercial purposes, such as salesmen behaviour (Yakasai & Jan, 2015) and job interview performance (Naim, Tanveer, Gildea, & Hoque, 2015).

The entertainment and gaming industry is increasingly interested in personality. The industry aims to understand and use human behaviour for consumer products, such as game recommender systems (H.-C. Yang & Huang, 2019), and building movie consumption profiles (Palomba, 2020).

The majority of personality studies, including its links to achievement and success, are in the field of education (Lounsbury, Sundstrom, Loveland, & Gibson, 2003; Meyer, Fleckenstein, Retelsdorf, & Köller, 2019; Demetriou, Kazi, Spanoudis, & Makris, 2019; Israel, Lüdtke, & Wagner, 2019). Some research focused on early recognition of students with troubling behaviour (Adnan, Mukhtar, & Naveed, 2012) or expected academic failure (Uddin & Lee, 2016).

Recent studies explored the relation between anger and personality traits, further pushing the boundaries between emotion and personality traits (Pease & Lewis, 2015). Kwiatkowska and Rogoza (2019) investigated the relation between modesty and personality traits. Eduardo and Ildefonso (2020) studied the link between personality traits and car crashes. Ābele, Haustein, Møller, and Zettler (2020) explored the relationship between anger and personality. The latter studies aimed to design better educational driving programs to reduce crashes, reduce and manage anger while driving, and enhance driving skills.

Medical and health-related research has recently highlighted the importance of personality traits in placebo and non-placebo medical trials, highlighting the relation between optimism and placebo response (Kern, Kramm, Witt, & Barth, 2020). Interestingly, the relation between anxiety and increased nocebo response. In mental-health studies, personality traits are used for planning health treatments (Bucher, Suzuki, & Samuel, 2019). Furthermore, personality traits can assist in conceptualizing treatment cases by highlighting strengths and identifying barriers and limitations. Surprisingly, personality traits were linked to some diseases, such as asthma (Najjab, Palka, & Brown, 2020). This has led to further research to identify and understand personality traits and unearth the psychophysiological connection.

1.1 Research Motivation

Human personality is very intriguing. The ability to understand and convey personality is becoming crucial in the age of user experience. Understanding and predicting human behaviour requires understanding human personality traits.

There is a massive increase in research targeted toward personality psychology in many areas requiring the development of automatic personality recognition systems. New technologies have re-sparked interest in HCI. Technology should be able to socially interact with people based on their personality. However, debate continues about the best model for personality recognition. Surprisingly, little attention has been paid to the effect of honest social signals on the big five personality traits and personality recognition. However, by understanding the process of personality recognition, it can be embedded in computers and robots to manifest personality and enhance the user experience.

Therefore, research on automatic personality recognition, existing corpora, and experimentation with the Speaker Personality Corpus (SPC) revealed major challenges. This has led to in-depth research on personality psychology, which has uncovered a massive gap between the two research communities, computer science and personality psychology. Accuracy of personality judgement research is still new and is the focus of much of the current research in psychology. In computer science, the research is built on datasets with incomplete personality evaluation. The research on accuracy and prediction has been more subjective, including judgements of either self or zero-acquaintances (strangers). The assumed ground truth underlying such subjective matter results in an incomplete personality ground truth score.

The most important element missing from personality datasets in the computer science community is knowledge of an acquaintance. Another important element is the use of the appropriate personality questionnaire and stimulus.

Computer scientists have been quickly collecting data and building applications to recognise personality; however, the applications are only as good as the data. Data serve as the solid base for the structure to be built. Data collection requires the knowledge of personality psychologists because, due to their knowledge and theories, computer scientists are capable of building more credible systems that can make it to commercial development and deployment in many areas.

The motivation of this research is to increase the accuracy of personality judgement by bridging the gap between the computer science community and personality psychology community. Research in personality recognition in the computer science community must go hand in hand with progress in the psychology community. Personality recognition research in the computer science community is moving without any strong relationship and guidelines from the psychology community. The motivation of the research is to strengthen this relationship by building a dataset for the computer science community based on the rules or models created and introduced in the psychology community. Moreover, guidelines for judgemental accuracy in the computer science community must be based on the guidelines used and well researched in personality psychology community.

There have been several suggestions for improving personality recognition in computer science; unfortunately, no studies have followed these recommendations (Vinciarelli & Mohammadi, 2014a; J. Joshi, Gunes, & Goecke, 2014; Vinciarelli et al., 2015; Lehmann-Willenbrock, Hung, & Keyton, 2017). Surprisingly the four previously referenced papers received between 22 and 68 references since their publication. This is clear evidence that the gap has not yet been fulfilled.

1.2 Aims and Objectives

The main aim of this research has been to design and validate a personality recognition model. The model is able to identify, capture, and analyse social signals from non-verbal acoustic cues. Social signals are used to identify the big five personality traits. The proposed model relies on the correlation between social signals and the big five personality traits. There are three initial primary aims of the research described in this thesis:

1. To investigate the relationship between non-verbal acoustic cues and honest social signals.
2. To determine the extent to which honest social signals correlate with the big five personality traits.
3. To develop a system which captures and analyses honest social signals through non-verbal cues to identify the big five personality traits and improve personality recognition.

However, to this end, the research reported in Chapter 6 of this thesis concluded that there were several limitations and in-appropriation of the SPC for personality recognition. The aims were adjusted to appropriately reflect the need to design a new corpus:

1. To design and build the Personality Traits Corpus.
2. To analyse the new corpus and perform exploratory data analysis.
3. To test the corpus against several machine learning algorithms.
4. To determine a reliable algorithm for personality recognition from non-verbal acoustic cues from speech.

1.3 Research Questions

There have been recent developments in automatic personality recognition. However, the research is still in its infancy compared to other mature computer science topics. This thesis raises several crucial questions related to automatic personality recognition and its development and growth.

The core question of this thesis is ***Can we recognise personality from non-verbal acoustic cues?*** Personality recognition from verbal cues, such as transcripts and text, has been the subject of a fair amount of research. In contrast, personality recognition from non-verbal acoustic cues has received limited attention.

The next question this work will explore is ***What are the acoustic–prosodic features that can reflect personality traits?*** Different feature extraction techniques were used on the Speaker Personality Corpus (SPC) and the new Personality Traits Corpus (PTC). The results are presented in Chapter 6 and Chapter 9.

Can we use machine learning algorithms to recognise the big five personality traits from non-verbal acoustic cues? This question is explored in three different settings. In the first setting, the SPC and machine learning algorithms are used to classify personality traits from non-verbal acoustic cues, which revealed several limitations. The second setting involved the use of a subset of SPC with machine learning algorithms. The results of both settings are discussed in Chapter 6. The third setting included the new Personality Traits Corpus and machine learning algorithms. The results are presented in Chapter 9.

Is it possible to build a corpus based on judgemental accuracy model from psychology? The Realistic Accuracy Model (RAM) is taken from the psychology community and used to build the Personality Traits Corpus (PTC). This is demonstrated in Chapter 8, the chapter describes the design and setting of the experiment, platforms, and applications.

What is the best machine learning algorithm for personality recognition from non-verbal acoustic cues from speech? This question is answered in Chapter 9. It was apparent from several experiments that κNN was the best-performing machine learning algorithm when used on the PTC. However, this conclusion would be more credible if more data were collected and experimented with. Chapter 9 presents these experiments and their results.

What traits can be recognised accurately from non-verbal acoustic cues from speech? Chapter 9 confirms that all personality traits can be recognised but with varying accuracies. It was concluded from these experiments that the hardest trait to recognise is conscientiousness. However, by expanding the corpus it is possible to improve the recognition results.

1.4 Major Contributions

The research described in this thesis presents original contributions in the areas of HCI and psychology. The major contributions are as follows:

- **Contribution 1:** Reviewed and experimented with available personality datasets. This has led to exposing a major gap between personality psychology and personality research in HCI.
- **Contribution 2:** Building a corpus for speaker personality traits from audio. The new Personality Traits Corpus was created with over four hours of speech. This corpus is the first in the computer science community to be built using the Realistic Accuracy Model (RAM). The new Personality Traits Corpus correctly calculated ground truth from self and acquaintances, and analysed the corpus. The corpus offers many possibilities for future research onto identifying personality traits from verbal and non-verbal cues.
- **Contribution 3:** Automatic personality traits recognition from non-verbal cues. Experimented with eighteen classifiers to identify the big five personality traits from speech using only acoustic–prosodic features.

1.5 Thesis Structure

The remainder of the thesis is organized as follows:

- **Chapter 2.** Personality. This chapter provides information about the definition of personality and trait theories. This chapter focuses on the history of the big five. In addition, it presents an overview of social signals and social signal processing in the literature.
- **Chapter 3.** Automatic Personality Recognition. This chapter presents a detailed review of research in the area of automatic personality recognition. It covers personality recognition from verbal and non-verbal cues.
- **Chapter 4.** Existing Personality Datasets. This chapter describes the available personality corpora. The chapter also highlights their limitations.
- **Chapter 5.** Experimental Methodology. This chapter describes the setting for the experiment with the SPC. It gives a brief description of all machine

algorithms. In addition, the chapter includes a brief explanation of the feature extraction techniques as well as evaluation metrics.

- **Chapter 6.** Experiments with the Speaker Personality Corpus. This chapter trains and builds several models in the experimental setting described in Chapter 5. It highlights the limitations of the SPC and again experiments with a subset of the SPC to overcome its limitations. The results of both experiments are discussed.

- **Chapter 7.** Accuracy of Personality Judgements. This chapter has a brief background about personality psychology. In addition, it explains the Life Story Interview and the RAM.

- **Chapter 8.** The Personality Traits Corpus (PTC). This chapter details the data collection process for the new PTC. It includes the ground truth process and exploratory data analysis.

- **Chapter 9.** Experiments with the Personality Traits Corpus. This chapter trains several machine learning algorithms and applies feature reduction techniques on the new corpus. Moreover, it explains and implements data augmentation and builds new classification models. Results for the PTC and the augmented PTC are discussed in the analysis section.

- **Chapter 10.** Conclusion and Future Work. This chapter highlights the main contributions of this thesis and identifies the limitations of the study. Finally, the chapter presents possible future research opportunities.

Chapter Summary

This chapter highlights the motivation of this research, its questions, and contributions. Additionally, it briefly discusses the chapters that follow.

Chapter 2

Personality

Until a character becomes a personality it cannot be believed.

Walt Disney

2.1 Definition of Personality

In psychology, there have been many definitions of personality. However, they all originate from Allport's (1961) definition: "*Personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristics behaviour and thought*". J. D. Mayer (2007) published an article presenting recent definitions of personality. Pervin, Cervone, and John (2005) slightly modify Allport's definition, adding the term 'consistent' to emphasize that personality is stable. Personality "*refers to those characteristics of the person that account for consistent patterns of feelings, thinking, and behaving*" (Pervin et al., 2005).

In the beginning of the 20th century, there was a rising interest in personality theories (John, Robins, & Pervin, 2010). There are several approaches to personality:

1. The Psychoanalytic Approach: Developed by Sigmund Freud. The theory divides the mind's processes and thoughts into conscious and unconscious behaviours (Westen, Gabbard, & Ortigo, 1990).

2. The Biological Approach: This approach argues that inherited genes have a substantial effect on human personality (Bouchard & Loehlin, 2001).

3. The Humanistic Approach: The approach emphasizes a human's responsibility to grow and learn and to be self-accepting of his life choices (Burger, 2014).

4. The Cognitive-Social Learning Approach: First proposed by Bandura and Walters (1977), it claims that behaviour affects how a human acts in a social setting, and the social setting has an effect on the human's behaviour. Later, he added the cognitive factor, which states that a human's cognitive abilities play an important role in his behaviour in a social setting and vice versa. This conclusion was reached by Walter Mischel (Allen, 2015).

5. The Trait Approach: Gordon Allport's approach focused on objectifying personality (Myers, 2004). Allport and Odbert (1936) described traits as consistent patterns of an individual's characteristic behaviour. The big five is a trait theory.

2.2 Personality Trait Theory

2.2.1 Gordon Allport Traits

Allport was a pioneer in the filed of trait theories (Carducci, 2009). Allport and Odbert (1936) argued that personality psychology should be an independent field of psychology. Allport's research has led to the trait theory. Allport and Odbert (1936) extracted 17,935 terms that reflect human behaviour from the unabridged English dictionary. This is known as the '*Lexical Approach*' (John, Angleitner, & Ostendorf, 1988). The authors discovered four categories: personality traits, temporary states of mood, characterial evaluations, and miscellaneous (physical characteristics). However, there is no clear boundary between the categories. Some terms overlapped between the categories.

2.2.2 Raymond Cattell's 16 Traits

Cattell (1966) built on the lexical approach of Allport and Odbert's list. Cattell used empirical clustering methods on a subset of 4500 traits, reducing the list to 35 variables. Cattell performed factor analysis and discovered 12 factors. In addition, Cattell (1966) created the 16 Personality Factors Questionnaire (16PF). He claimed his 16PF showed excellent results with self-review and observer review. However, its technicality and focus on universal personality in groups led to a loss of interest in the psychology community.

2.2.3 Hans Eysenck's PEN Theory

Eysenck used factor analysis to propose three personality dimensions: extroversion, neuroticism, and psychoticism (PEN) (Eysenck, 1950). Although he used factor analysis and similar techniques to Cattell, his approach yielded different results. Moreover, Eysenck is one of the few theorists who used experimental methods to test his hypotheses (Allen, 2015).

2.2.4 Tupes, Christal, and Norman's First Five Factors

In 1961, Tupes and Christal performed a study on recurrent personality factors. The authors used Cattell's 35 traits and factor analysis. This produced five factors: surgency, agreeableness, dependability, emotional stability, and culture. Unfortunately, this study was published as a technical report (Wiggins, 1996). This affected its popularity in the research area of personality psychology.

W. T. Norman (1963) replicated Allport and Odbert's lexical approach methods. Norman extracted 18,125 traits from an English dictionary. However, the terms were classified into seven categories: stable traits, temporary states, activities, social roles, social effects, evaluative terms, anatomical and physical terms, and ambiguous terms which are unrelated to personality. Norman applied factor analysis to the traits category. He replicated the five factor model of Tupes and Christal (1961).

2.2.5 Lewis Goldberg's "The Big Five"

After Norman's discovery, there was a slight inactivity in trait and personality theories. However, research by Goldberg (1981) renewed interest in research on personality traits. Goldberg used Norman's list to test the generalizability of the five factors. Goldberg (1981) argued that the five factors should be oblique bipolar dimensions. He performed several studies and concluded that factor analysis produced the same five factors (Goldberg, 1981, 1990, 1992; Goldberg & Saucier, 1998). Therefore, he coined the term '*Big Five*'.

2.2.6 Costa and McCrae's NEO Inventories

In 1987, McCrae and Costa (1987) claimed the validity of the big five. The authors used self-reports and peer ratings to compare the adjectives of the five factor model with their NEO (neuroticism, extraversion, and openness to experience) personality questionnaire. Their first model used cluster analysis on Cattell's (1966) 16PF. However, they redefined and extended their model to include agreeableness and conscientiousness. McCrae and Costa (1995) published papers providing evidence of the big five retrieved from questionnaires (McCrae & John, 1992).

The NEO personality questionnaire originally contained 240 items (Costa & McCrae, 1992a). Due to its length, Costa and McCrae (1992b) reduced the questionnaire to 60 items. Further research by Benet-Martinez and John (1998) created the big five inventory (BFI). BFI-44 indicates 44 questionnaire items. In a published study, Rammstedt and John (2007) investigated the reliability and validity of the BFI-10. Unfortunately, it only captured only 70% of the BFI-44, although it maintained 85% retest reliability. Therefore, BFI-10 is recommended if research time is limited.

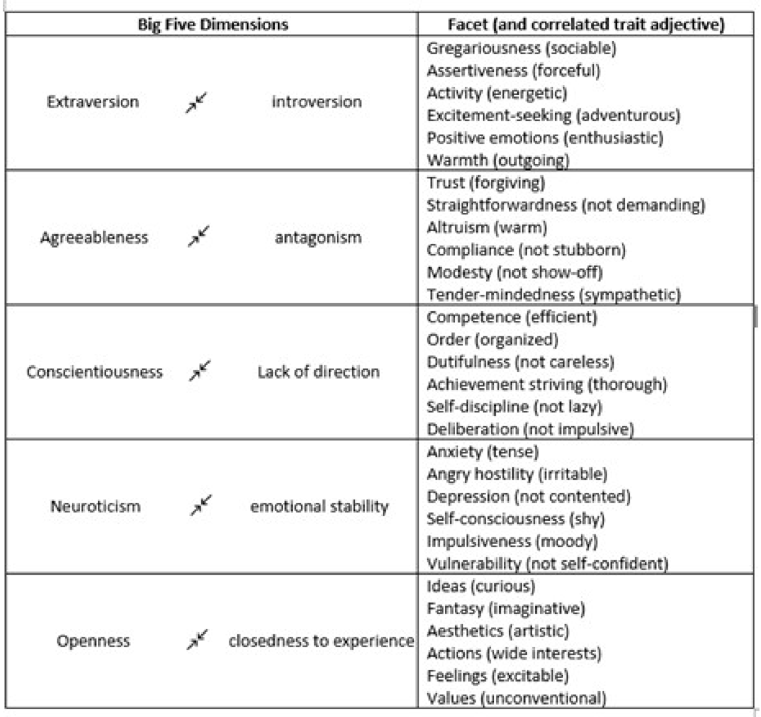
2.2.7 The Big Five

The big five have been generated through two different approaches: a lexical approach (Allport & Odbert, 1936; Goldberg, 1981) and personality inventory questionnaires (Eysenck, 1950; McCrae & Costa, 1987). The big five personality traits are five dimensions, each including six facets (Goldberg, 1993). Figure 2.1 shows the five factors and their related facets.

The big five can be described as follows (Costa & McCrae, 1992b; Goldberg, 1981; McCrae & Costa, 1987; McCrae & John, 1992):

1. Openness: People who attain a high measure of openness tend to be imaginative, curious, and artistic. It also indicates interest in a wide range of activities and topics. However, people with low scores are more simple and traditional. Low scores also indicate limited interests and less curiosity.

2. Conscientiousness: People with high conscientiousness show organization, reliability, and self-dependence. Conscientious people plan ahead and are hard-working. In contrast, low measured people are characterised by laziness, carelessness, and a lack of ambition.



Big Five Dimensions	Facet (and correlated trait adjective)
Extraversion ↗↙ introversion	Gregariousness (sociable) Assertiveness (forceful) Activity (energetic) Excitement-seeking (adventurous) Positive emotions (enthusiastic) Warmth (outgoing)
Agreeableness ↗↙ antagonism	Trust (forgiving) Straightforwardness (not demanding) Altruism (warm) Compliance (not stubborn) Modesty (not show-off) Tender-mindedness (sympathetic)
Conscientiousness ↗↙ Lack of direction	Competence (efficient) Order (organized) Dutifulness (not careless) Achievement striving (thorough) Self-discipline (not lazy) Deliberation (not impulsive)
Neuroticism ↗↙ emotional stability	Anxiety (tense) Angry hostility (irritable) Depression (not contented) Self-consciousness (shy) Impulsiveness (moody) Vulnerability (not self-confident)
Openness ↗↙ closedness to experience	Ideas (curious) Fantasy (imaginative) Aesthetics (artistic) Actions (wide interests) Feelings (excitable) Values (unconventional)

Figure 2.1: *The big five factors and their relative facets (John & Srivastava, 1999)*

3. Extraversion: Extroverts are sociable, friendly, and outgoing. Meanwhile, introverts are perceived as timid, quiet, and task oriented.

4. Agreeableness: People with high agreeableness are appreciative, sympathetic, and forgiving. In addition, they are flexible and open-minded. On the contrary, people measuring low in agreeableness can be stingy, uncooperative, and suspicious.

5. Neuroticism: Unlike other factors, people who score high for neuroticism exhibit self-pity, anxiousness, and insecurity. However, people who score low present calmness, satisfaction and patience.

The existing literature on the big five is extensive and focused on its in-depth research and applications (Matz, Chan, & Kosinski, 2016). The big five are stable across instruments, observers, and self-reports (McCrae & Costa, 1987). Moreover, the big five model is the most accepted model for personality recognition (McAdams & Pals, 2006; Pianesi, 2013). This further supports the ability of the big five dimensions to represent an individual's differentiating characteristics. A large and growing body of literature has recognised the advantages of the big five in academia (Komarraju & Karau, 2005; Lounsbury et al., 2003; O'Connor & Paunonen, 2007), businesses (Dong, Lepri, & Pentland, 2012; Gilal, Jaafar, Basri, Omar, & Tunio, 2015; Ridgell & Lounsbury, 2004; Yakasai & Jan, 2015), and health (Booth-Kewley & Vickers, 1994; Halama & Gurnáková, 2014; Jerram & Coleman, 1999).

There are several ways to recognise personality traits (Isbister & Nass, 2000). For example, extroverts can be identified from text when using words representing affirmation, such as ‘definitely’ (Pennebaker & King, 1999). Moreover, when a person is resting his hands next to his side he may be an introvert (Cassell, 2000b). Conversely, talkative and excited people are identified as extroverts (Mairesse, Walker, Mehl, & Moore, 2007). Some personality markers are as follows:

1. Verbal: linguistic features are words, their attributes (letter, tense, types), and emotions (positive, negative). Pennebaker and King (1999) argued that personality and individual differences appear in words. Further, the authors demonstrated personality stability across time and regardless of topic.

2. Non-verbal: divided into prosodic and visual features. Schötz (2002) described prosodic features as attributes related to voice, such as pitch and volume. Meanwhile, Gavrilescu (2015b) showed that visual features are related to facial expressions (gaze, eyebrow, and lip movements) and body gestures (hand gestures and body posture).

2.3 Social Signal Processing

In recent years, there has been an increased amount of literature on ‘social signals’ (Pantic & Vinciarelli, 2014). Pentland (2007b) coined the term ‘social signals’, and Poggi and Francesca (2010) defined it as follows: “*A social signal is a communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions, social emotions, social attitudes and social relationships*”. In 2010, Pentland and Heibeck stressed that certain unconscious signals can predict a social outcome, such as in speed dating (Pentland, 2007a). The authors named them ‘honest signals’. These signals are as follows (Pentland & Heibeck, 2010):

- Influence: this signal focuses on the influence one speaker has on another. It measures the extent to which a speaking pattern of one speaker affects another.

- Mimicry: measures the copied gestures and expressions between two speakers, such as smiles and nodding.

- Activity: high activity in a conversation is an indication of interest and excitement.

- Consistency: measures the coherence and cohesion of voice attributes and gestures in a conversation. Unsteady and unstable gestures or a sudden change of volume or pitch indicates a lack of mental focus and a higher possibility of being influenced by others.

Previous research by social psychologists such as Allport (1937) have established that using thin slices of behaviour observation does not affect prediction accuracy. This was confirmed by Ambady and Rosenthal (1992) through experimental studies and effect size analyses. There is a growing number of published studies demonstrating the positive effects of applying social signals in predicting human behavioural outcomes in different social environments (Ambady & Rosenthal, 1993).

Bousmalis, Mehu, and Pantic (2009) demonstrated the use of social signals to detect an agreement or disagreement in a social situation. Their method divided non-verbal cues into agreement and disagreement cues. Further, they provided testing data and tools for interested researchers to prove their hypotheses. Riggio and Feldman's (2005) book included several applications in health, business, and education where non-verbal communication is present in social interactions. The authors stressed the benefits of understanding such cues. For example, in healthcare, non-verbal behaviour can be related to states of mental or physical health. In education, Harrigan, Rosenthal, Scherer, and Scherer (2008) explained the importance of understanding non-verbal behaviour to the learning process and student/teacher relations. Feldman (2014) highlighted the benefits of recognizing and displaying non-verbal cues in several business settings, such as advertising, marketing, sales, job interview, negotiations, management, and leadership.

2.3.1 State of the Art

Understanding human behaviour in HCI has led to growth in research on personality and behaviour analysis (Pantic et al., 2011; Pantic & Vinciarelli, 2014; Vinciarelli, Pantic, Bourlard, & Pentland, 2008a). Current research aims to find variables related to the recognition of various social signals (Brunet, Donnan, McKeown, Douglas-Cowie, & Cowie, 2009). Brunet et al. argue that the most important social signals are those which appear unconsciously, which they refer to as 'honest signals'. Moreover, their research highlights the main problems associated with social signal processing: social signal preprocessing and social signal analysis. Much of the current literature on social signals pays particular attention to preprocessing (capturing non-verbal communication).

A number of studies attempted signal analysis, which includes extracting and interpreting the data (Batrinca, Lepri, Mana, & Pianesi, 2012b; Cristani, Raghavendra, Del Bue, & Murino, 2013; Gatica-Perez, Vinciarelli, & Odobez, 2014; Vinciarelli, Salamin, Polychroniou, Mohammadi, & Origlia, 2012). Some signal analysis research has focused on speech and prosodic features only (Mohammadi, Vinciarelli, & Mortillaro, 2010; Mohammadi & Vinciarelli, 2012; Ranganath, Jurafsky, & McFarland, 2013).

An interesting paper by Vinciarelli et al. (2008b) categorised non-verbal behaviour cues into five codes (Figure 2.2). A code or a combination of codes are responsible for a certain function, such as deceiving and detecting deception.

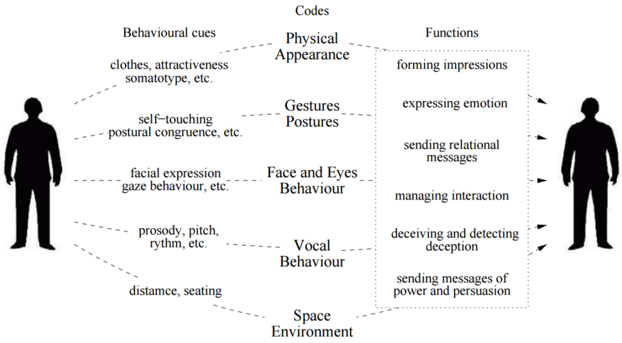

Codes
Behavioural cues
Functions
Physical Appearance
clothes, attractiveness somatotype, etc.
forming impressions
Gestures Postures
self–touching postural congruence, etc.
expressing emotion
sending relational messages
facial expression gaze behaviour, etc.
Face and Eyes Behaviour
managing interaction
prosody, pitch, rythm, etc.
Vocal Behaviour
deceiving and detecting deception
sending messages of power and persuasion
distamce, seating
Space Environment

Figure 2.2: *Non-verbal cues classified into codes and fulfilling different functions (Vinciarelli et al., 2008b) - Used with permission*

The subsequent sections present the latest research related to social signal extraction approaches.

2.3.1.1 Physical Appearance

There is a lack of research on social signal extraction from appearance, such as age, clothing, body shape, skin, and hair color (Vinciarelli et al., 2008a). Most research focused on perceiving age or beauty and attractiveness (Aarabi, Hughes, Mohajer, & Emami, 2001). For example, Aarabi et al. (2001) used a genetic algorithm (Melanie, 1996) to learn how people rate photographs. Similarly, Sutić, Brešković, Huić, and Jukić (2010) focused on beauty by learning how people rate photographs. The study used κ-nearest neighbour (κNN) (Dudani, 1976), Adaboost (Freund, 1995), and neural networks (Lippmann, 1987). However, their study produced low classification scores.

Kalayci, Ekenel, and Gunes (2014) presented an experimental study on attractiveness and beauty from videos, from which they extracted static and dynamic features. Data training was performed using support vector machine(SVM) (Cortes & Vapnik, 1995) and random forests (Breiman, 2001). Another experimental study by Eisenthal, Dror, and Ruppin (2006) used κNN and SVM. Their results were similar to human raters.

Aghaei, Parezzan, Dimiccoli, Radeva, and Cristani (2017) studied the relation between clothing styles and social interaction. Similarly dressed people tend to have a longer social interaction, while those dressed differently may only exchange non-verbal greetings. This is still an under-researched topic in literature.

2.3.1.2 Body Gestures and Posture

A thorough examination of the relevant literature revealed that the way in which people recognise social signals from body gestures has received little attention. McKeown, Curran, McLoughlin, Griffin, and Bianchi-Berthouze (2013) collected data on body movement during different types of laughter. The database is available for researchers exploring laughter generation scenarios during human–robot interactions. In an investigation of laughter based on body movement, Griffin et al. (2013) found that gestures and movements were related to different laughter types. The authors reported that their models for automatic recognition of laughter performed well. Similarly, Varni, Camurri, Coletta, and Volpe (2009) added a phase synchronisation feature to body movements to recognise empathy and dominance-related signals.

In another study, Gaschler et al. (2012) showed the importance of body gestures and head poses in successfully recognising social behaviours in human–robot interaction. An experimental study by Y.-C. Yu (2016) proved the benefits of a dual feedback system in an e-learning environment. Capturing head movements assisted instructors in understanding their students and enhancing their engagement.

Roudposhti, Nunes, and Dias (2015) used social signals to understand the relation between body movements and interpersonal behaviour and estimate the social role of the individual (Leader).

2.3.1.3 Face and Eye Behaviour

One of the most important studies on face and eye behaviour was conducted by Ekman and Friesen (1978). Their research introduced the Facial Activity Coding

System (FACS), which measures all observable facial movements. It uses action units (AU) to describe different facial activities. Jiang, Valstar, and Pantic (2012) proposed a system for better capturing social signal based on real-time AU and improved detection system. The human face conveys many expressions. Using facial expressions, researchers have been able to recognise a set of preset emotions, such as fear, anger, and frustration (J. Chen, Chang, & Tu, 2015). Al-Samarraie, Eldenfria, and Dawoud (2017) explored the connection between information-seeking behaviour and the big five by tracking eye movements. The results showed that the difference in the speed of information seeking is based on certain personality traits.

Martinez, Valstar, Jiang, and Pantic (2017) reviewed the recent advances in FACS, highlighting its limitations and challenges. The authors suggested guidelines for future research using FACS.

2.3.1.4 Vocal Behaviour

Research in social signal interpretation from vocal features is extremely limited. Moreover, studies have focussed on speech recognition rather than vocal features (Vinciarelli et al., 2008a). Further research by Vinciarelli, Valente, Yella, and Sapru (2011) used prosodic social signals to identify different roles of speakers in a meeting. The authors compared the formal role of the speaker to the identified social role. A recent study by Kim, Valente, Filippone, and Vinciarelli (2014) used a regression approach with extracted prosodic features to predict conflict outcome in political debates. Their recommendation for future studies was to include more non-verbal cues, such as facial expressions. Feng et al. (2020) utilised vocal social signals to recognise depression through the use of convolutional neural networks.

An interesting study proposed the use of mobile phones to monitor vocal social signals to recognise stress and reschedule meetings in the user's calendar if stress levels were high (Pejovic & Musolesi, 2015).

In INTERSPEECH 2013, a paralinguistic challenge was announced (Schuller et al., 2013). The aim was the recognition of social signals from the 'SSPNet Vocalization Corpus'. The classification labels were laughter, fillers, and garbage. Several papers were submitted (An, Brizan, & Rosenberg, 2013; Bone et al., 2013; Gosztolya, Busa-Fekete, & Tóth, 2013; Janicki, 2013; Kirchhoff, Liu, & Bilmes, 2013; Krikke & Truong, 2013; Oh, Cho, & Slaney, 2013; Wagner, Lingenfelser, & André, 2013). The winners were Gupta, Audhkhasi, Lee, and Narayanan (2013), who successfully enhanced performance using time-series smoothing and masking.

Following the INTERSPEECH 2013 challenge, Brueckner and Schuller (2013) and Brueckner and Schulter (2014) performed different classification techniques on the corpus. The results surpassed those of the winning paper. H. Joshi, Verma, and Mishra (2020) experimented with the corpus and produced better results using deep long short-term memory (LSTM).

Flutura, Wagner, Lingenfelser, Seiderer, and André (2016) proposed an interesting idea, collecting and analysing social signal cues through the use of mobile phones. The authors experimented in both controlled and uncontrolled environments. Not surprisingly, accuracy fell in the uncontrolled environment setting during data collection and when testing.

Nasir, Baucom, Georgiou, and Narayanan (2017) tested whether acoustic features could predict marital therapy outcomes. Their results demonstrated that the performance was better than expert-coded behaviour.

2.3.1.5 Space Environment

Space environments refer to the distance between individuals and where they stand in a socially interacted environment (Vinciarelli et al., 2008b). C. W. Chen, Wu, and Aghajan (2011) examined the social interaction environment. Their aim was to predict social interaction between two people and to better understand social group dynamics. The authors studied body movements and distance between all participants. Unfortunately, they have not yet published their results. Cristani, Murino, and Vinciarelli (2010) proposed several possible research agendas involving the use of environment and intelligent surveillance. Zhu, Li, Zhao, and Jiang (2018) studied the relation between personality traits and different scenes, such as arenas, jail cell, valleys, and the sky. The study used linear regression to train the model, and which successfully recognised personality traits from user-liked pictures.

2.3.1.6 Multi-non-verbal Cues

Vinciarelli, Dielmann, Favre, and Salamin (2009) built a very rich political debate corpus known as Canal 9. The authors intention was to study the social interaction between speakers. The corpus was annotated with turn-taking, agreement, and disagreement. Kim, Filippone, Valente, and Vinciarelli (2012) experimented on the topic of politics further by building their own political corpus and studying conflict perception from these clips.

Okwechime, Ong, Gilbert, and Bowden (2011) proposed mining social signals. The authors used a combination of prosodic features, gaze, and body movement successfully predicting conversational interest. Similarly, Moreno (2012) proposed interpreting social signals from gazes, gestures, and body movements to recognise important behaviours in children on playgrounds, such as aggressiveness and depression.

In an educational setting, Jang, Lee, Kim, and Cho (2013) aimed to identify the primary social signals associated with student engagement and confirmation in a 1:1 student–teacher setting. Their non-verbal communication included facial expressions, gestures, and posture. While their results were encouraging, further research is required for generalization.

Bousmalis, Mehu, and Pantic (2013) surveyed different datasets and proposed several methods to identify the social cues associated with the detection of agreement and disagreement.

Gatica-Perez (2014) built their own corpus for a real temporary job interview and extracted non-verbal features, such as gestures and vocals. The authors stressed the importance of social signals in the workplace, for example, in interviews and team building.

Griffin et al. (2015) focused on a single behaviour, laughter. Their research showed all possible non-verbal cues associated with laughter (gestures and facial expressions). The aim of the research was to recognise the different cues associated with different laughter types, such as real, fake, and awkward laughter. The automatic laughter recognition scores were relatively similar to observer rating scores. The authors claimed their findings could lead to the use of laughter in human–robot interaction (HRI).

Leone, Migliorisi, and Sessa (2016) found social signals that relate to honesty and deception through a combination of facial expressions and gestures. Moreover, they confirmed the benefits of using a multi-non-verbal modal compared to a single modal to capture non-verbal cues.

Social cues were also used to study the rapport between virtual agents and humans (Cerekovic, Aran, & Gatica-Perez, 2016). Interestingly, paralinguistic cues and turn-taking were found to be correlated with self-reported rapport.

Navarathna, Carr, Lucey, and Matthews (2017) experimented with predicting movie ratings based on facial expressions and body movements while people are watching a movie. Facial expressions were used to determine whether or not a person was engaged with a movie on the screen.

Kasano, Muramatsu, Matsufuji, Sato-Shimokawara, and Yamaguchi (2019) classified individuals' confidence based on the use of social cues. The authors concluded that vocal features and head motions can be an indication of a person's level of confidence.

2.3.2 Applications

Social signals appear in any human interaction within a social context. Understanding human behaviour in a social context is essential in different fields, such as business (Ambady, Krabbenhoft, & Hogan, 2006; Chattopadhyay, Dahl, Ritchie, & Shahin, 2003), education (D'Errico, Leone, & Poggi, 2010; Lepper, Woolverton, Mumme, & Gurtner, 1993), and healthcare (Aufegger, Bicknell, Soane, Ashrafian, & Darzi, 2019; Tanaka et al., 2017; Tapus & Mataric, 2008). Moreover, HCI and HRI will benefit from recognising and successfully interpreting social signals. Humans tend to interact with computers as they do with humans (Nass & Brave, 2007). Indeed, Nass, Steuer, and Tauber (1994) declared that "*Computers are social actors (CASA)*". Further, Nass highlighted the similarities between a human's reaction in human–human interaction and HCI scenarios (Nass & Moon, 2000).

The social factor changes the HCI experience (Vinciarelli, Pantic, et al., 2012). HCI is interactive and requires participation from both sides (Salah, Pantic, & Vinciarelli, 2011). Accordingly, the research is greatly shifting toward social interaction. Currently, HCI design include the social factor (Esposito, Esposito, & Vogel, 2015).

The CASA paradigm applies to HCI and HRI (Lee, Peng, Jin, & Yan, 2006). Researchers have long shown interest in robot development and intelligence. Recently, researchers have begun to focus on social robots (Aly, Tapus, et al., 2012; Miwa, Umetsu, Takanishi, & Takanobu, 2001; Woods, Dautenhahn, Kaouri, Boekhorst, & Koay, 2005). Robots with social interaction abilities understand different human behaviours and needs (Syrdal, Koay, Walters, & Dautenhahn, 2007).

2.3.3 Challenges

Social signal processing aims to enable computers to recognise and understand social signals in HCI and HRI. Recently, a considerable stream of literature has grown up around social signal processing. However, social signal processing has several challenges (Pentland, 2007b; Vinciarelli et al., 2008b, 2008a; Vinciarelli, Salamin, & Pantic, 2009). These main challenges are as follows:

1. Social signals require multimodal approaches to capture the different non-verbal cues (Kim et al., 2014; Varni et al., 2009).
2. Controlled settings constrain social behaviours, and people are aware of the experiments. Thus, the study may not yield the same results if implemented in a real-world setting.
3. The use of real-world data delivers a more realistic measure of the effectiveness of technological modelling.

Chapter Summary

To date, a number of studies have confirmed the effectiveness of the big five in personality recognition (Celli, Pianesi, Stillwell, & Kosinski, 2013), academic achievements (Lounsbury et al., 2003), health studies (Halama & Gurnáková, 2014), and career success (Ridgell & Lounsbury, 2004). Furthermore, these studies (McAdams & Pals, 2006; Pianesi, 2013) highlighted the stability and reliability of the big five. Considering all of this evidence, this research will adopt the big five as a measure of personality.

This chapter also introduced the definition of social signals, state of the art social signal processing research, and its challenges. Social signal processing is still in the early stages of research (Vinciarelli et al., 2008b). Much of the research has focused on extracting behavioural cues to recognise social signals relating to a function, such as laughter or detecting deception. However, this thesis aims to extract behavioural cues to identify non-verbal social signals that identify personality traits.

Chapter 3

Automatic Personality Recognition

> *Your smile is your logo, your personality is your business card.*
>
> Jay Danzie

HCI is a morphing field (Carroll, 1997). It adapts to changes presented by new technologies (Lazar et al., 2010). Rogers (2004) noted that HCI is focused on designing a better interface for the user. However, people behave differently toward technology based on their individual characteristics (Sigurdsson, 1991). Now, HCI involves the design of an experience for people with different behavioural characteristics and attributes, such as race, age, gender, personality, needs, and abilities (Rogers, 2004). Cassell (2000a) argued for the need for personality recognition and application in embodied conversational agents (ECA). For example, tutoring could be customised to personality and progress rather than just based on student progress (Komarraju & Karau, 2005; Uddin & Lee, 2016).

3.1 Background

In the 1970s, user cognitive behaviour was first experimented with by R. Mayer (1975). The authors explored the possibility of designing a cognitive model of programmer behaviour. The aim was to simplify programming for non-programmers. Although a model was proposed, the experiment was performed in a controlled environment, and thus the model must be verified. There was no clear indication of programmer behaviour, as the authors reported that some students benefited from flowcharts, some were hindered, and some showed no difference. R. Mayer (1975) claimed that an understanding of programmer experience in each language and semantic knowledge leads to better design of computer languages. However, each user is an individual with different characteristics. Regardless of programming language, the interaction should adjust to individual abilities, needs, and characteristics. Mayer's (1975) research had major significance to human factors researchers (Carroll, 1997). Designing systems with better usability requires human factors groups to collaborate with computer scientists and software engineers (Carroll, 1997; Shneiderman, 1981).

Much of the literature since the end of the 1970s has emphasised usability and user interface design (Carey, 1997; Galitz, 2002, 2007; Martin & Eastman, 1996; Nielsen, 1989; Shneiderman, 1979, 1983, 1998; Shneiderman & Plaisant, 2003; Shneiderman, Plaisant, Cohen, & Jacobs, 2017). Research focuses have included menu location, colour schemes, input options, screen layout, and navigation.

In recent years, the literature on HCI became concerned with personality and the big five personality traits (Celli et al., 2013; Vinciarelli & Mohammadi, 2014b; Wright & McCarthy, 2008). Most literature about big five personality traits recognition was largely based on empirical studies investigating the best combination of features and algorithms for recognising personality traits (Celli, Lepri, et al., 2014; Schuller et al., 2015). The following sections review the literature on personality recognition through verbal cues, non-verbal cues, and non-social signals.

3.2 Non-verbal Cues

Personality recognition based on non-verbal cues includes the following aspects: speech attributes (except words), prosody, gestures, facial expressions, and body movements (Vinciarelli & Mohammadi, 2014b). Polzehl, Moller, and Metze (2010) focused on non-verbal speech features, such as prosody, spectral features, and Mel-frequency cepstral coefficients (MFCC) (Davis & Mermelstein, 1980). Using SVM, they reported good results of around 60% for all traits.

Staiano, Lepri, Subramanian, Sebe, and Pianesi (2011) emphasized on low-level features (LLF) and high-level features (HLF) of personality traits. The authors focused on several algorithms, including naïve Bayes (N. Friedman, Geiger, & Goldszmidt, 1997), hidden Markov models (HMM) (Rabiner & Juang, 1986), and SVM to evaluate the effectiveness of the features in terms of identifying the personality traits. Extraversion scored best with HMM at 73.1% and neuroticism highest at 63.9% with naïve Bayes.

A study by Chastagnol and Devillers (2012) suggested using the sequential floating forward search algorithm (SFFS) (Pudil, Novovičová, & Kittler, 1994) on speech, which consists of an alternating forward and backward search. The forward search adds feature enhancing performance, and backward search removes LLF. The authors used a greedy version of the algorithm, which stops the iteration when higher performance is achieved. Although their training results performed above the baseline, their test results failed to exceed the baseline except for openness and agreeableness. Moreover, this approach lacks generalisability.

Batrinca, Lepri, and Pianesi (2011) highlighted the benefits of using a multi-modal recognition system of acoustic and visual cues. The authors chose three machine learning algorithms: naïve Bayes, SVM with a linear kernel, and SVM with a radial basis function (RBF) (Cortes & Vapnik, 1995). The results were high for extraversion, conscientiousness, and neuroticism. Extraversion achieved the highest score using a single acoustic feature. Similarly, Batrinca, Mana, Lepri, Pianesi, and Sebe (2011) investigated personality traits through self-introduction videos. The authors used naïve Bayes and SVM on non-verbal acoustic and visual cues. The authors reported high scores for extraversion and conscientiousness. Meanwhile, Mairesse et al. (2007) used naïve Bayes on prosodic features only. The lowest accuracy was 50% for the agreeableness trait.

In 2012, Valente, Kim, and Motlicek (2012b) studied an Augmented Multiparty Interaction (AMI) meeting corpus[1]. The authors implemented boostexter (Schapire & Singer, 2000) on linguistic and non-linguistic features separately. The study showed better results for non-linguistic features for all traits. In a study investigating prosodic features, Mohammadi and Vinciarelli (2012) reported enhanced results using SVM with a Gaussian kernel (Cortes & Vapnik, 1995). Extraversion and conscientiousness scored above 70%, while the remaining traits scored above 60%. Similarly, Mohammadi, Origlia, Filippone, and Vinciarelli (2012) performed a study on the same dataset but selected a different approach for feature extraction. However, they reproduced relatively similar results to the previous research.

Audhkhasi, Metallinou, Li, and Narayanan (2012) conducted a study on 640 speech clips. Three techniques were proposed, including Gaussian mixture models (Reynolds, 2015), within-class covariance normalization (WCCN) (Hatch, Kajarekar, & Stolcke, 2006), and a tree-structured Bayesian network (N. Friedman et al., 1997). Their research focused on prosodic features only. WCCN and the tree-structured Bayesian network performed better in recognising personality traits than Gaussian mixture models.

[1] https://groups.inf.ed.ac.uk/ami/corpus/

In 2012, Batrinca et al. (2012b) experimented with personality recognition through task collaboration in HCI environment. Different levels of collaboration resulted in the recognition of different personality traits. Extraversion and neuroticism outperformed the remaining traits regardless of collaboration level. In a follow-up study, Batrinca, Mana, Lepri, Sebe, and Pianesi (2016) compared results of automatic personality recognition through task collaboration in HCI and human-human interaction (HHI) scenarios. A multi-modal system of acoustic and visual cues was proposed. The results showed that extraversion and neuroticism are easily recognised in HCI.

Wagner, Lingenfelser, and André (2012) proposed segmenting the training set and extracting LLF from meaningful segments and applied k-means clustering (Hartigan & Wong, 1979). Inhomogeneous clusters will have their LLF pruned, and unrelated frames will be ignored. At the end of the process, HLF are extracted from the pruned clusters and used to train the SVM classifier. The development set performed slightly better than the baseline. However, the test failed to show significant improvement, especially for openness and extraversion. This might have been caused by the clustering and pruning approach, as important features may have been pruned.

Lepri et al. (2012) experimented with only a single personality trait, extraversion. The focus was on connecting the trait with the behaviour in a small-group meeting environment. The feature extraction combined audio features (e.g. speaking time) and eye gazes. Their findings showed that audio features and eye gazes are ineffective for classifying the trait if not combined together. However, the gazes of others or social attention from others toward a silent target produced statistically significant performance.

Alam and Riccardi (2013) investigated automatic personality recognition from two different corpora: broadcast news and conversation. The authors only extracted acoustic features. Feature selection was performed using a combination of information gain (Y. Yang & Pedersen, 1997) and relief (Kononenko, 1994) with sequential minimum optimization (SMO) (Platt, 1998), random forest (Breiman, 2001), and Adaboost (Freund, 1995). Conscientiousness and extraversion scored the highest among traits, while the remaining traits scored in the 60% range.

Salamin, Polychroniou, and Vinciarelli (2013) proposed a double experiment to recognise personality traits and conflict in mobile phone conversations. The authors used the SSPNET - Nokia Corpus (Polychroniou, Salamin, & Vinciarelli, 2014). The features that were focused on were head movements and acoustic features, separately and combined. The only trait that performed slightly better than the baseline was neuroticism using only acoustic features. This could have been the result of the corpus type, which was a conflict corpus, and could be clearly apparent in acoustic features if the user is calm or tense, thus triggering neuroticism.

A small-scale study by Gavrilescu (2015a) focused on personality recognition through facial expressions. The study used neural networks (Lippmann, 1987) and successfully recognised extraversion, neuroticism, and openness. However, the small amount of training data may have resulted in lower scores for conscientiousness and agreeableness.

Pohjalainen, Räsänen, and Kadioglu (2015) explored different feature reduction techniques solely and combined on the Speaker Personality Corpus (SPC) and with acoustic features extracted from openSMILE (Eyben, Wöllmer, & Schuller, 2010). The authors concluded that each feature set was unique to a personality trait, and no features or feature sets were found generic across all personality traits. Moreover, their trait average of 64% with a reduced feature set was similar to the average when all features were included.

Jothilakshmi and Brindha (2016) experimented with non-verbal cues differently than most studies. The authors suggested that spectral structures convey highly essential linguistic features. Frequency domain linear prediction (FDLP) is a parametric description of speech temporal dynamics. It is calculated by applying the discrete cosine transform (DCT) and then performing linear prediction on the DCT output. The results showed that κNN outperformed SVM and multi-layer perceptron (MLP) (Rosenblatt, 1958).

Aydin, Kindiroglu, Aran, and Akarun (2016) participated in the Chalearn Lap 2016 First Impressions Challenge (Escalante et al., 2016; Ponce-López et al., 2016a). The dataset consisted of videos extracted from the YouTube website. The authors extracted audio and visual features. The visual features were facial features and motion energy. The regressor they used was random forest. Interestingly, the feature sets had the same performance levels across all traits. When feature sets were combined, the performance slightly increased. Ventura, Masip, and Lapedriza (2017) used the same dataset to extract visual and audio features and used deep learning as the regression of choice. The authors concluded that visual features alone produced better results than audio features and combined audio and visual features. Moreover, Gürpınar, Kaya, and Salah (2016) extracted facial, audio, and ambient features to feed into a convolutional neural network (CNN). Their proposed method which achieved a mean score of 91.3%, ranked first in the First Impressions Challenge.

A very interesting paper by Carbonneau, Granger, Attabi, and Gagnon (2017) investigated audio non-verbal cues differently than in previous research. The paper explained the transformation of an audio track to a spectrogram. Then, the authors trained and built a model based on the spectrogram patch images. Surprisingly, the results yielded a level of performance similar to that of the state of the art techniques presented in other research. This could be an indication of a future perspective for personality recognition that needs further research.

Multi-modal and bi-modal systems are very common for personality recognition. More papers are focused on multi-modal and bi-modal personality recognition than on mono-modal or unimodal systems. K. Yang, Mall, and Glaser (2017) proposed bi-modal deep learning recognition of personality from short first-impressions videos. Visual and audio features were extracted, and their proposed models showed that visual features are superior to audio features. However, superiority does not imply that visual features are facial features alone and may or may not include different visual elements, such as the background, clothes, and body. Their results placed them in the top five competitors in the ChaLearn Challenge (Ponce-López et al., 2016b).

Gilpin, Olson, and Alrashed (2018) extracted non-verbal cues from the SPC. The extracted features were MFCC, pitch, and energy. The authors used the full dataset for training and building the classifier. The test set was a new dataset collected by the authors and followed a similar protocol to the SPC. The authors used SVM and HMM. Their experimentation demonstrated the highest accuracy for conscientiousness and agreeableness.

Cai et al. (2018) examined physiological changes to analyse the correlation between personality traits and emotions using wearable devices. The authors captured facial cues, such as winks, and body gestures, such as a clenched fist. One of their findings was the link between winks and agreeableness. Another finding was that emotions have an influence on the relationship between traits and behaviour.

Hoppe, Loetscher, Morey, and Bulling (2018) sought to recognise personality from eye gazes using a device mounted on the participant. The authors collected data from 42 participants running errands at the university. Their classifier successfully predicted all personality traits except openness. However, the authors highlighted that they only collected a small dataset.

Suen, Hung, and Lin (2019) collected their own data and built a model. The dataset was a collection of audio-video interviews; however, only facial features were extracted from each frame. The model was built using convolutional neural network

(CNN). Their average accuracy was 95%, exceeding that in previous personality research.

Beyan, Zunino, Shahid, and Murino (2019) proposed a novel approach to extract visual activity from the key dynamic non-verbal features of images. The authors tested their proposed approach on two different multi-modal datasets. Their results demonstrated that the novel approach succeeded with the Emergent LEAder corpus (ELEA-AV) dataset (Sanchez-Cortes, Aran, & Gatica-Perez, 2011), with an average accuracy of 72%. However, their approach was not more accurate with ChaLearn dataset compared to other published results.

Table 3.1 summarizes the recent research on automatic personality recognition from non-verbal communication.

3.3 Verbal Cues

There is a relatively small body of literature that is concerned with personality recognition through verbal cues. (Celli et al., 2013). However, it is slowly growing and expanding.

In 2007, Mairesse et al. (2007) published a pioneering study on personality recognition from dialogue. Speech was recorded using an Electronically Activated Recorder (EAR) (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). Linguistic and prosodic features were combined with several techniques, such as SVM, naïve Bayes, Adaboost, and C4.5 (Quinlan, 2014). The results revealed successful prediction of personality traits using naïve Bayes for all traits except for conscientiousness, which was best predicted with SVM.

Ivanov et al. (2011) studied personality recognition from HHI. Their study highlighted the use of linguistic and prosodic features from speech. The proposed system consisted of openSMILE-based feature extraction (Eyben et al., 2010) and the boostexter classifier (Schapire & Singer, 2000). Their results were significant for conscientiousness and extroversion. Neuroticism achieved the lowest accuracy of 32.8%.

In 2013, Biel, Tsiminaki, Dines, and Gatica-Perez (2013) examined speech activity and facial expressions of emotion. The authors chosen algorithms were random forests and linguistic inquiry and word count (LIWC), developed by Pennebaker, Francis, and Booth (2001). The proposed model achieved the highest scores for agreeableness, followed by conscientiousness and neuroticism. The worst performing trait was openness.

Gievska and Koroveshovski (2014) followed up on Biel et al.'s (2013) recommendation to extend the feature set. The authors used the same dataset and incorporated emotions and gender as well as audio and visual features. They experimented with different levels of emotions while maintaining lexical and visual features. The best feature combination was audio, visual, emotions, emotion valence, and frequency. Their average $F1$ score was 71%.

Poria, Gelbukh, Agarwal, Cambria, and Howard (2013) recommended using sentiment analysis and lexical features. Their method was based on opinion mining using sentiments. Their proposed model achieved an average accuracy of 63.6% for personality traits recognition.

Detailed examination of personality recognition using a multi-modal system of verbal and non-verbal cues by Sarkar, Bhatia, Agarwal, and Li (2014) showed that non-verbal features performed better for extraversion, conscientiousness, and neuroticism. Alam and Riccardi (2014b) proposed a similar multi-modal model which combined verbal and non-verbal cues. Non-verbal cues performed better than verbal cues as a single feature option. However, the multi-modal model reported better results.

Verhoeven, Soler Company, and Daelemans (2014) discussed their submission for the Workshop on Computational Personality Recognition in 2014. The challenge presented participants with a text-based dataset. The authors used token unigrams, character trigrams, linguistic inquiry word count (LIWC), and Soler 2014. Soler 2014 is a feature set with different types: character-based, word-based, dictionary-based, and syntactic features. The authors experimented with features solely and combined. Their support vector classifier (SVC), regardless of feature or combined features, did not yield outstanding performance. In fact, the authors recommended further research on personality recognition and feature parameter tuning.

Another study by Alam and Riccardi (2014a) explored different combinations of verbal and non-verbal cues using two datasets. The three verbal features were tokens, part-of-speech, and LIWC. Non-verbal features were acoustic features. On both datasets, LIWC performed better than other verbal features. However, acoustic features outperformed all verbal features.

An et al. (2016) used their own corpus to recognise personality from speech. The authors used different feature sets separately and combined. Their feature sets were low-level descriptors (LLD), LIWC, dictionary of affective features, and fundamental frequency variances features. The best feature or feature set was personality-trait dependent. Their mean for unweighted average recall was 40.62%.

Majumder, Poria, Gelbukh, and Cambria (2017) used deep learning to recognise personality traits from text. The authors used Mairesse et al.'s (2007) baseline features set to extract document-level features. The authors also filtered sentences and discarded those without emotion-related words. Next, they extracted word-level features using word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and n-grams. Finally, they trained with CNN and classified with MLP. The results showed a slightly improved performance compared to the baseline of Mairesse et al. (2007).

Text can be a rich context for personality recognition. Tandera et al. (2017) experimented with two datasets consisting of Facebook users' statuses: myPersonality (Stillwell & Kosinski, 2004) and a dataset collected manually from Facebook. The authors compared the performance of classic algorithms to that of a deep learning technique. The results showed enhanced accuracy for all traits, exceeding that of previous studies.

With the introduction of deep learning, more research articles targeting personality have used deep learning. For example, J. Yu and Markov (2017) investigated personality recognition from Facebook users' status updates and applied three different deep learning approaches: fully connected, CNN, and recurrent neural network (RNN). All three approaches delivered similar results, with a mean of 60%. Similar research by Yuan, Wu, Li, and Wang (2018) used the same dataset and LIWC as feature extraction and deep learning for trait recognition. The best accuracy was achieved for openness (76%).

An and Levitan (2018) used their speech deception corpus with native English and Chinese speakers to experiment with homogeneity. Linguistic and prosodic features were extracted from speech. Eight models were built along two heterogeneous dimensions: native language and gender. The homogenous results were better than the non-homogenous results.

An interesting multi-modal study tested several verbal and non-verbal features on a video corpus (Aslan & Güdükbay, 2019). Four cues were explored: ambient features, such as clothes, lighting, and distance to objects, as well as facial features, audio prosodic features, and text transcriptions. The authors stated that the accuracy of their model exceeded that of previous reports. It was concluded that facial features gave the most promising results, followed by ambient and audio–prosodic features. The worst performance was with text-based features.

Rissola, Bahrainian, and Crestani (2019) experimented with the EAR dataset (Mehl, Gosling, & Pennebaker, 2006) and used a capsule-based model. This type of model groups the similar properties of an entity together. The authors showed an increase in performance when compared with the LIWC baseline.

In a recent paper, Han, Huang, and Tang (2020) built a personality recognition model to recognise personality from Chinese users microblogs. Their model was based on the correlation between personality traits and semantic word categories.

Zhao, Zeng, Xiao, Che, and Wang (2020) proposed using a long short-term memory (LSTM) model with user text sentiment and profile as features. The features were converted to attention information, which was was fed into the LSTM to predict the users' personality. Their attention-based classification model achieved an average $F1$ score of 72%.

Table 3.2 summarises recent research on automatic personality recognition from verbal communication.

3.4 Non-Social Signals

Personality recognition has become an intriguing topic for many researchers. The focus has shifted to personality recognition from any user-generated content. Such content can be user profiles on social media, user behaviour on a media platform, handwriting, or portraits.

One of the first studies on personality recognition based on non-social signals was conducted by Chittaranjan, Blom, and Gatica-Perez (2013). Their aim was to link personality recognition to mobile phone usage behaviour. Collected data included SMS logs, application usage, call logs, and Bluetooth usage. The authors outlined the features, which correlated with the big five traits. However, their classifier mean $F1$ score was only at 57%.

In 2015, Youyou, Kosinski, and Stillwell (2015) compared the personality recognition accuracy of computers and friends. The authors employed users' likes as a behaviour for personality recognition and compared whether a human friend or a computer provided an answer closer to a self-rated personality questionnaire. Their research concluded that computers were better judges of personality than humans.

Another study on Facebook's user profile images (Segalin et al., 2017) found that warm-coloured images reflected extroversion and agreeableness traits, while indoor images were linked to neuroticism. The authors also concluded that computer classification of personality traits was more accurate than that of human judges.

Wei et al. (2017) built a complex experiment centred around users' heterogeneous information ensemble (HIE) collected from Weibo, which records users' tweets. The collected information included tweets, avatars, emoticons, and responsive behaviour. Each type of data collected had a feature extraction layer and a classifier layer. The personality score from each of the classifiers was input into the stacked ensemble algorithm, which learned how to best combine predictions from different well-performing machine learning algorithms. The authors stated that their HEI model performed very well and achieved higher results than state of the art models for all personality traits.

Recently, there has been growing interest in personality recognition from physiological cues. Wache et al. (2015) collected electrocardiogram (ECG), galvanic skin response (GSR), facial-electroencephalogram (EEG), and facial emotional responses. Their preliminary experiment showed high score for the openness trait using GSR features. The rest were below the chance level. Klados et al. (2020) demonstrated high accuracies for all personality traits from EEGs alone.

A very interesting study provided a different perspective on personality recognition from different user-generated content. In their research, Taber and Whittaker (2018) studied the difference in the personality perception of users between offline, Snapchat, and Facebook. On Facebook, users seemed less open, less aggregable, and less neurotic than their offline personality. In their second study, the authors compared Snapchat, Facebook, and offline user personality perceptions. Snapchat showed a more extroverted and open personality perception than both Facebook and offline. It seemed that social anxiety and people's or followers' judgement has a great impact on users when using Snapchat.

Moreno-Armendáriz, Martínez, Calvo, and Moreno-Sotelo (2020) proposed personality recognition from portraits originating from YouTube videos. The research adopted deep neural networks (DNN) for automatic feature extraction and classification. The authors reported an average accuracy of 65.86%. Similarly, Kachur et al. (2020) used static facial images from photographs and artificial neural networks to predict the big five personality traits. The authors stated that their results were better than state of the art methods reported in the literature.

In 2018, Kaushal and Patwardhan (2018) published an extensive survey on personality recognition from social media. The media platforms covered were Facebook and Twitter. Different features were used for both social media platforms, including likes, profile images, followers, friends, and retweets. Furthermore, Azucar, Marengo, and Settanni (2018) conducted a meta-analysis of recent studies on individuals' digital foot print on social media. The authors identified the scarcity of personality research as a major limitation. Another study by Ganguli, Mehta, and Sen (2020) focused on surveying the recent machine learning algorithms used in personality recognition from social media. The authors discussed three different types of social media sets: Twitter, Facebook, and Linkedin.

Humans are a rich source of information, and human behaviours are linked to personality. Therefore, Kamisaka and Ishikawa (2020) studied customers' behaviour during shopping. Their aim was to capture visual data from customers and build a shopping profile based on their personality. Their ground truth was based on online and hybrid (use online and shop offline) customers because they had already built a log of their shopping behaviour and completed a self-rated personality questionnaire. Information from offline-only customers served as their test data. The authors reported excellent accuracy for personality prediction of offline customers based on their visual cues, which were collected from the store using activity and motion sensors.

Table 3.1: *Comparing the results of research papers on automatic personality recognition through non-verbal communication*

Ref.	Dataset Type/Size	Dataset Name	Features	Algorithm/Technique	Measure	Trait				
						O	C	E	A	N
(Mairesse et al. 2007)	Conversation Script/96	NA	Prosodic features	Naïve Bayes (NB)	Accuracy	64.56%	52.11%	67.56%	50.44%	61.78%
(Polzehl et al., 2010)	Audio / 30 clips Participants (1 participant)	NA	Prosody, spectral features, Mel Frequency Cepstral Coefficients (MFCC).	Support Vector Machine (SVM)	Accuracy	60%				
(Batrinca, Lepri, & Pianesi, 2011) (Batrinca, Mana, et al., 2011)	Self-presentations/89	NA	Prosodic features Visual cues	Support Vector Machine Radial Basis Function(SVM-RBF) Support Vector Machine Linear (SVM- Lin) Naïve Bayes (NB)	Accuracy	66.29% (SVM-Lin)	73.03% (SVM-RBF)	70.78% (SVM-RBF)	65.16% (SVM-RBF)+ (SVM-Lin)	76.40% (SVM-RBF)
(Staiano et al., 2011)	Video Meetings/4	NA	Activity level Emphasis Speaking features Prosodic features Social attention	Support Vector Machine Radial Basis Function (SVM-RBF) Support Vector Machine Linear (SVM- Lin) Naïve Bayes (NB) Hidden Markov Model (HMM)	Accuracy	54.7% (SVM-RBF)	58.3% (NB)	73.1% (HMM)	58.3% (NB) (SVM-RBF)	63.9% (NB)
(Mohammadi & Vinciarelli, 2012)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Prosodic features	Logistic Regression (LR) Support Vector Machine with Gaussian Kernel (SVM)	Accuracy	60.1% (SVM)	72.5% (LR)	73.5% (SVM)	63.1% (SVM)	66.1% (LR)
(Mohammadi et al., 2012)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Prosodic features Voice quality features	Ordinal Regression (N Categories)	Accuracy	63.9% (N=3)	70.8% (N=3)	78.6% (N=3)	65.8% (N=3)	72.0% (N=3)
(Audhkhasi et al., 2012)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Prosodic features	GMM WCCN (hatch2006within) Tree-structured Bayesian Network (N. Friedman et al., 1997)	Accuracy	67.57 WCCN	74.29 BN	74.29 BN	74.29 BN	71.17 WCCN
(Chastagnol & Devillers, 2012)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Prosody spectral features	SFFS	Accuracy	58.0%	75.5%	73.4%	65.0%	65.2%
(Wagner et al., 2012)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	OpenSMILE Prosodic features	Support Vector Machine (SVM)	Accuracy	60.1%	77.6%	79.2%	56.1%	63.0%
(Batrinca et al., 2012a) (Batrinca et al., 2016)	Video Task/43	NA	Prosodic features visual cues	Support Vector Machine (SVM) with linear kernel	Accuracy	60.46%	69.76%	81.39%	69.77%	81.30%
(Lepri et al., 2012)	Video\12 meetings \4 people \6 hours\ 48 participants	Mission survival task corpus	Speaking activity Gaze direction	SVM with RBF kernel	Accuracy	X	X	59.95%	X	X
(Alam & Riccardi, 2013)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	openSMILE acoustic	Random Forest SMO Adaboost	UAR	65.2%	75.3%	83.0%	66.0%	69.2%
(Salamin et al., 2013)	60 mobile phone calls (120 subjects in total) 60 calls \11 hours and 48 minutes	SSPNet Mobile Corpus	Speech features (praat) pitch stylization model syllabus head movement	Support Vector Machine (SVM) with Radial Basis Function (RBF)	Accuracy	60.7% (only speech features)	53.3% (only movement)	59.2% (speech and movement)	61.7% (speech and movement)	62.5% (only speech features)
(Gavrilescu, 2015a)	Video\64 participants watching video	NA	Facial expressions Visual cues	Feed-Forward Neural Network	Accuracy	76.2%	52%	80%	50%	84%
(Pohjalainen et al., 2015)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	openSMILE Prosodic features	κNN	UAR	59.5%	79.6%	76.1%	59.2%	65.6%
(Jothilakshmi & Brindha, 2016)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Frequency Domain Linear Prediction features acoustic	Support Vector Machine (SVM) with polynomial kernel, multi-layer perceptron (MLP) κNN	Accuracy	92.66% κNN	95.22% κNN	98.06% κNN	97.83% κNN	98.37% κNN
(Aydin et al., 2016)	Video\10000 clips	First Impressions V2	Audio features Visual features	Random Forest	Accuracy	90.197%				
(Gürpınar et al., 2016) (Gürpinar, Kaya, & Salah, 2016)	Video\10000 clips	First impressions (ECCV '16, ICPR '16)	Audio (openSMILE) Facial expression scene	Random Forest	Accuracy	91.69%	91.66%	92.06%	91.61%	91.49%
(Ventura et al., 2017)	Video\10000 clips	First impressions (ECCV '16, ICPR '16)	Facial features	Convolutional Neural Network	Accuracy	91%	91.4%	91.5%	91.2%	90.7%
(Carbonneau et al., 2017)	Speech /640 clips	SSPNet Speaker Personality Corpus (SPC)	Audio spectrograms	Support Vector Machine (SVM)	UAR	56.3%	68.3%	75.2%	64.9%	70.8%
(K. Yang et al., 2017)	Video\10000 clips	First impressions (ECCV '16, ICPR '16)	Audio cues Visual cues	Bimodal-LSTM (L1, L2)	Accuracy	91.36% L2	92.2% L2	91.1% L2	89.77% L1	90.33% L1
(Gilpin et al., 2018)	Speech /640 clips + their own test corpus	SSPNet Speaker Personality Corpus (SPC)	Kaldi speech recognition toolkit Acoustic features	Support Vector Machine (SVM) Hidden Markov Model (HMM)	Accuracy	78.83% SVM	93.75% HMM	70.16% SVM	74.22% HMM	82.81% HMM
(Hoppe et al., 2018)	42 subjects had eye tracker walk around campus for 10 minutes	NA	Eyegaze	Random Forest	F1	30.8%	43.1%	48.6%	45.9%	40.3%
(Suen et al., 2019)	120 subject video for job interview\20 minutes each	NA	Visual features Facial detection	CNN	Accuracy	97.4%	96.7%	97%	90.9%	94.8%
(Beyan et al., 2019)	Video\10000 clips	First impressions (ECCV '16, ICPR '16)	Visual cues	Support Vector Machine (SVM)	AUC	77%	80%	81%	79%	79%
(Beyan et al., 2019)	Video\27 meetings with 3 or 4 participants\15 minutes average	ELEA-AV	Visual cues	Support Vector Machine (SVM)	Accuracy	69%	70%	77%	79%	67%

Table 3.2: *Comparing the results of research papers on automatic personality recognition through verbal communication.*

Ref.	Dataset Type/Size	Dataset Name	Features	Algorithm/Technique	Measure	Trait				
						O	C	E	A	N
(Mairesse et al., 2007)	Conversation Script/96	NA	LIWC MRC Prosodic features	Naïve Bayes (NB) AdaboostM1 (ADA) Sequential Minimal Optimization (SMO) J48 Decision Trees (J48) Nearest Neighbor (NN) JRip Rules Set (JRIP)	Accuracy	57% (NB)	67.67% (NB)	73.0% (ADA)+(NB)	61.33% (NB)	73.89% (NB)
(Ivanov et al., 2011)	24 speakers/Conversation scripts/119	NA	openSMILE linguistic Prosodic features	Boostexter	Accuracy	40.34%	94.96%	63.03%	56.30%	30.77%
(Valente et al., 2012b)	Meeting Sessions/ 32	NA	Speech activity features Prosodic features N-gram Words Dialogue act tags Participants interaction features	Boostexter	Accuracy	57.1%	67.6%	74.5%	55.4%	68.7%
(Biel et al., 2013)	Vlog script/442	NA	LIWC N-grams	Support Vector Machine (SVM) Random Forests (RF)	R2	.05 (RF)	.10 (RF)	.07 (RF)	.18 (RF)	.10 (RF)
(Poria et al., 2013)	Essays\2400	NA	LIWC MRC Sentic Based Emotional Features Bag of concepts Sentic feature Negation	Sequential Minimal Optimization (SMO)	F1	66.1%	63.3%	63.4%	61.5%	63.7%
(Alam & Riccardi, 2014b)	Vlog script/404	YouTube Personality Corpus	Audio-Video (AV) Lexical (Lex) LIWC POS Emotional	Sequential Minimal Optimization (SMO)	F1	65.0%	61.9%	71.0%	76.7%	61.4%
(Gievska & Koroveshovski, 2014)	Vlog script/404	YouTube Personality Corpus	Audio cues Visual cues Tokenization, part-of-speech (POS) tagging, root words extraction, and stemming Emotion related words	Support Vector Machine (SVM)	UAR	71.4%	51.8%	76.8%	78.6%	73.2%
(Sarkar et al., 2014)	Vlog script/404	YouTube Personality Corpus	Audiovisual cues Sentiments Word statistics Gender Text	Logistic Regression	UAR	71.4%	76.8%	71.4%	78.6%	60.7%
(Verhoeven et al., 2014)	Vlog script/404	YouTube Personality Corpus	Token unigrams Character Trigrams LIWC Soler2014	Support Vector Machine Classifier (SVC)	F1	53.2%	59.6%	59.6%	57.4%	63.4%
(Alam & Riccardi, 2014a)	Speech Clips/640	SSPNet Speaker Personality Corpus (SPC)	Acoustic features Tokenization(bag of words) Parts of speech pos LIWC	Sequential Minimal Optimization (SMO)	UAR	63.1%	79.6%	78.2%	65.1%	66.9%
(An et al., 2016)	172 subject pairs\23 hours of speech	NA	OpenSMILE LIWC Affect features Fundamental frequency features	Sequential Minimal Optimization (SMO)	UAR	52.5%	43.4%	39.7%	45.3%	47.9%
(Majumder et al., 2017)	2,467 anonymous essays	NA	Document features (Mairesse Featrues) Sentence filtering (NRC emotion lexicon) Word2vec	Support Vector Machine (SVM) Multi-layer Perceptron (MLP)	Accuracy	62.68% (MLP)	57.3% (MLP- Fully Connected)	58.09% (MLP)	56.71% (MLP)	59.38% (MLP)
(Tandera et al., 2017)	250 participants \10000 status	myPersonality	LIWC SPLICE SNA users friendship network Word embedding (glove)	Naïve Bayes Support Vector Machine Logistic Regression Gradient Boosting Linear Discriminant Analysis MLP LSTM GRU CNN	Accuracy	79.31% (MLP) Undersampling applied	62% (GRU)	78.95% (MLP) Undersampling applied	67.39% (CNN) Undersampling applied	79.49% (MLP) Undersampling applied
(J. Yu & Markov, 2017)	250 participants \10000 status	myPersonality	skip-gram uni-gram bi-gram tri-gram Author Information	Deep Neural Network Fully connected architecture	F1	53.1% All	51.2% All	63.3% Author Information	55.8% Author Information	55.6% All
(Yuan et al., 2018)	250 participants \10000 status	myPersonality	LIWC	Convolutional Neural Network	Accuracy	76%	58%	57%	57%	60%
(An & Levitan, 2018)	125 hours of speech in the corpus from 173 subject pairs and 346 individual speakers.	NA	LIWC OpenSMILE Dictionary of Affect in Language (DAL) Word vectors Part of speech	Support Vector Machine	Accuracy	Kindly refer to the paper because the breakdown of accuracy too complex to include in the table.				
(Aslan & Güdükbay, 2019)	Video\10000 clips	First Impressions V2	Facial features Audio Environment Text	LSTM	Accuracy	91.66%	92.14%	92.08%	91.89%	91.62%
(Rissola et al., 2019)	Conversation Script/96	NA	Semantic features Utterance Tokens Facial features	Capsule Neural Network	F1	60.87%	75%	60.87%	66.67%	60%
(Han et al., 2020)	Weibo microblogs \409 subjects	NA	Lexicon LIWC Bag of Words Latent Semantic Analysis and Second Order Attribute	Support Vector Machine Logistic Regression Random Forest	F1	74.3%	72%	70.8%	71.5%	69.2%
(Zhao et al., 2020)	250 participants \10000 status	myPersonality	Word2vec Sentiment analysis	LSTM	F1 Recall	72.2% 65.78%				

Chapter Summary

Automatic personality recognition can be based on verbal communication, non-verbal communication, and text. Social signals cannot be interpreted from text. Therefore, personality recognition from text is outside the scope of this research. Thus, this chapter examined the state of the art research related to automatic personality recognition through verbal and non-verbal communication. Most research on automatic personality recognition has been experimental. Research groups have used small datasets and applied different feature selection methods and machine learning algorithms. Thus far, only a few studies have presented multi-modal structures for personality recognition. It was apparent from the previous research that there were varying results and extracted features. This variation may be linked to several factors, such as the use of different datasets, audio extraction tools, feature extraction techniques, machine learning algorithms, and evaluation criteria. Consequently, the results of different models cannot be compared. Automatic personality recognition is still a new area of research, and so the introduction of social signals may have a positive effect. This research aims to improve automatic personality recognition by using social signals.

This chapter also covered some advances in personality recognition from non-social signal cues, such as portraits and handwriting.

Chapter 4

Existing Personality Datasets

An investment in knowledge always pays the best interest.

Benjamin Franklin

Personality recognition requires large datasets or repositories (Finnerty, Lepri, & Pianesi, 2016; Mohammadi & Vinciarelli, 2012). The large volume of published literature repeatedly uses the same corpora (e.g. Biel & Gatica-Perez, 2013; Mohammadi et al., 2010). The subsequent section briefly describes several of the available corpora.

4.1 Speaker Personality Corpus (SPC)

The SPC is a speech corpus is proposed by Mohammadi et al. (2010). There are 640 clips recorded with 330 different identities. The same identity is not present in both the training and test sets. A total of 309 clips represent journalists, and 331 represent non-journalists. Clips were extracted from 96 news bulletins from Radio Suisse Romande in 2005. The clips are in French and include a single speaker. The total length of the corpus is seven hours; the average length of each clip is 40 seconds, but the authors extracted 10 seconds from each clip. Therefore, almost all clips are 10 seconds long and were assessed by 11 judges using the BFI-10 questionnaire (Rammstedt & John, 2007). The judges were random strangers who volunteered to participate in the study, and they were not psychologists or personality experts. The judges had no knowledge of the French language, and thus the assessments were based on non-verbal cues. Each clip produced five dimensions reflecting the big five traits. For each judge, the scores for each trait for all clips were averaged, and each trait was scored high or low based on each judge's average score for each trait. Next, each clip was labelled as high or low for a certain trait based the on judges' majority agreement. In this corpus, when six or more judges agreed on a label for a single clip, the clip was labelled based on the majority vote.

4.2 SSPNet Mobile Corpus

Polychroniou et al. (2014) collected mobile data to support social signal processing research. Their corpus consisted of a collection of sixty telephone calls between 120 unacquainted individuals recruited from the University of Glasgow. The call included a task requiring the two subjects to negotiate a common solution. After the call was concluded, the subjects were asked to complete the BFI-10 questionnaire along with two additional questionnaires related to conflict and interpersonal attraction.

Several behavioural cues were annotated: speaking activity, laughter, overlapping speech, back channel, fillers, and silence. Unfortunately, due to their grant expiring, the SSPNet website no longer exists.

4.3 YouTube Personality Corpus

The YouTube dataset was collected by crowdsourcing personality impressions and audio-visual behavioural analyses from video logs on YouTube (Biel & Gatica-Perez, 2013). The preliminary data were collected in 2009, consisting of 2269 hours of video. Each video was 1 to 6 minutes long, and there was a total of 469 different users video logging. The collection process was restricted by choosing videos with the words 'vlogging'or 'vlog'. Moreover, only a single speaker appeared in each video, talking directly to the camera and showing only the area from the shoulders and above. Biel and Gatica-Perez (2013) explained the difficulties associated with annotating long hours of video and references the widely suggested 'thin slices'(Ambady & Rosenthal, 1992) as an alternative. The authors final set included 442 one-minute video logs, of which 47% were from males and 53% from females. Amazon's Mechanical Turk (MTurk) was used to crowdsource personality impressions. MTurk users were required to answer a questionnaire after watching the one-minute video log. Moreover, the dataset included audio and visual cues extracted from the one-minute video logs.

Audio cues included speaking activity, and prosodic cues. Visual cues included looking activity and pose.

4.4 Sociometric Badge Corpus

The concept of aociometric badges was introduced by Olguin, Paradiso, and Pentland (2006). Sociometric badges "*are wearable electronic badges capable of automatically measuring the amount of face-to-face interaction, conversational time, prosodic style, physical proximity to other people, and physical activity levels, using social signals derived from vocal features, body motion, and relative location*" (D. O. Olguín, 2007). The badge was used in an organisational setting in Chicago. A total of 1900 hours of data were collected from 23 out of 28 employees at a data server firm in one month. The collected data included the following (D. Olguín et al., 2009):

1. Employee performance to tasks: assigning time, closing time, assigned to, closed by, difficulty level, and number of follow-ups.

2. Employee behaviour: location relative to other employees or key locations he visited, such as printer and warehouse. It also records employee posture and activity data.

3. Interpersonal interactions: infrared (IR) is used when two employees are in face-to-face communication. The microphone records audio intensity.

The tasks were computer configuration tasks assigned to users on a first-come-first-serve basis. Each task had one of three difficulty levels. Employees submitted a configuration report and returned to the end of the queue.

4.5 ChaLearn First Impression V2

This is a very large dataset with 10000 videos built by Ponce-López et al. (2016a). Each video clip is 15 seconds long. Every clip has a single speaker who was looking at the camera and speaking in English about a random topic. It was extracted from YouTube. Each YouTube video was sampled up to six times, taking different segments from the same video. In addition, each YouTube channel was sampled up to three times, with up to three videos from each channel used for sampling. There were 3060 original videos and 2764 original channels. For annotation, every two clips were paired and presented to an Amazon's MTurk rater. Big five questionnaires were not used, only a variable representing each of the big five personality traits. For each clip, the ground truth is a value within the range [0,1]. For classification, the value is transformed using a threshold of 0.5. Thus, if the value is above the threshold the trait is high, and otherwise it is low.

4.6 Multimodal Human–Human–Robot Interactions (MHHRI)

Celiktutan, Skordos, and Gunes (2019) introduced a new dataset which included 18 subjects with 48 hours of interaction. There were two types of interaction: HHI and HRI. The human participants completed BFI-10 questionnaires about themselves and received acquaintance-assessed questionnaires. Each participant had nine to 12 acquaintance ratings. In the HHI, each participant was given a set of questions to ask their partner. Six out of eight questions were about robots, and there were two personal questions about a good and bad memory. For ground truth, the authors decided to use the acquaintances' scores only because there was higher agreement between them than between acquaintances and self-reports.

Table 4.1 summarizes the most popular corpora used for personality recognition and uses the big five as its personality measure.

Table 4.1: Popular big five personality trait corpora

Corpus	Ref	Data type	Number of Instances	Big Five	Ground Truth	Note
SSPNet Speaker Personality Corpus	Mohammadi et al. (2010)	Audio	640	BFI-10 questionnaire	Other-report (No acquaintance)	Single speaker per clip
SSPNet Mobile Corpus	Polychroniou et al. (2014)	Audio	120 subjects 710 minutes (60 calls)	BFI-10 questionnaire	Self-report	No longer available for download
YouTube Personality Corpus	Biel and Gatica-Perez (2013)	Video\audio	404 (28 hours)	Ten-Item Personality Inventory (TIPI)	Other-report (No acquaintance)	Single speaker per clip No longer available for download
Sociometric Badge Corpus	Olguin et al. (2006)	Audio	23 subjects (1900 hours)	Ten-Item Personality Inventory (TIPI)	Self-report	Big5 no longer associated with corpus
Chalearn V2	Ponce-López et al. (2016a)	Video\audio	10000 clips	5 variables (1 per trait)	Self-report	Single speaker per clip
MHHHRI	Celiktutan et al. (2019)	Audio	14 subjects (48 hours)	BFI-10 item questionnaire	Other-report (acquaintance)	Multiple speakers per clip

Chapter Summary

Although the Chalearn and Sociometric datasets contain the largest and richest data, unfortunately they cannot be used. The sociometric badge corpus does not have any data associated with the big five. Moreover, after further research, the sociometric badge is no longer available, as the creators decided to use it as a commercial tool for improving organisational roles and productivity. Regarding the Chalearn dataset, it has several issues that render it unfit for personality recognition. The first major issue is that a single variable is used to represent the big five trait instead of using a questionnaire. Second, Amazon's Mturk raters cannot assign the same level of a variable trait to both videos. This is because they must choose a video where a variable trait is more apparent. Lastly, the ground truth is crowdsourced from strangers, and there is no self-report or acquaintance report.

MHHRI is a well-built dataset; however, it has several limitations. The first limitation is the disregard of self-reports and sole reliance on acquaintance reports. Second, the experiment was done in the same order and repeated several times for each target and acquaintance to create more data and different pairs. However, this repetition has an affect on the acquaintances, as they have to repeat the same questions and answers several times to different targets. They become aware of the questions, which has an effect on authenticity and emotion since the repetition make it appear to be an emotionless response or answer.

The SSPNet Social Mobile Corpus, unfortunately, is one of many corpora that no longer exists or are no longer publicly available due to funding or grants expiring. Therefore, the only dataset available for experimentation is the SPC.

Chapter 5

Experimental Methodology

> *Decide what you want, decide what you are willing to exchange for it. Establish your priorities and go to work.*
>
> H. L. Hunt

This chapter introduces the experimental setting that will be used in the following chapters. It presents a brief description of the machine learning algorithms, acoustic feature extraction tool, feature reduction techniques, and evaluation measures that will be referenced throughout the remainder of this thesis.

5.1 Machine Learning Algorithms

This section briefly describes the functionality of machine learning algorithms and their advantages and drawbacks.

5.1.1 κ-Nearest Neighbors (κNN)

First introduced by (Dudani, 1976) in 1976. κNN is a memory-based algorithm (Blalock, 2003). By keeping the training data in its memory, it does not require fitting. κNN classifies instances according to majority vote based on the nearest-to-training instance. The κ distance can be calculated using four common distance measures: Euclidean, Hamming, Manhattan, and Minkowski distances. Euclidean distance is the distance between two vectors calculated as the square root of the sum of the squared differences between the two vectors. Hamming distance is the number of different bit positions between two equal-length binary strings. Manhattan distance is the distance between two points on a grid-like plane. It is also known as the city block difference or taxi-cab geometry. Minkowski is a generalization of the Euclidean and Manhattan distance measures. It has a parameter ρ which allows for choosing between Euclidean or Manhattan distance measures.

Since κNN is memory based, a large amount of training data is an advantage (Cunningham & Delany, 2020). In contrast, a high-dimensional dataset adds an extra layer of complexity in calculating the distance of neighbours and is time consuming (Gan & Gromiha, 2010). Another drawback is its incompetence when dealing with imbalanced datasets. κNN is biased toward the majority class because it will have more votes. In addition, outliers and noisy data affect prediction accuracy (W. Liu & Chawla, 2011).

5.1.2 Support Vector Classifier (SVC)

SVM (Cortes & Vapnik, 1995) can be used for regression and classification. SVC is used for linear and non-linear problems. The main idea is to create a hyperplane to separate the data into classes (Pedregosa et al., 2011). The algorithm accepts the training samples as input points and creates a line (or hyperplane) which separates the two classes. SVM selects the closest points, known as support vectors, to the line from each class. The distance between the support vectors and the separating line is called a margin. The goal of SVC is to maximize the margin for higher accuracy. For non-linear problems, the kernel parameters are tuned to change the data dimension and then create a hyperplane to separate the classes. The hyperplane must be as wide as possible to clearly separate the two classes. A major advantage is its effectiveness with high-dimensional datasets (Statnikov, Wang, & Aliferis, 2008). It performs best when there is a clear separation between classes. Unfortunately, the classes in the experimental corpus are not easily separable, and due to dimensionality, it requires a long training time (Westreich, Lessler, & Funk, 2010).

5.1.3 Random Forest (RF)

Random forest algorithm creates several decision trees with randomly selected data (Breiman, 2001). Each tree produces a prediction. The algorithm selects the best prediction through voting. RF has several advantages due to the number of decision trees created. A major advantage is its robustness to overfitting because decision trees average the predictions and thereby cancels any bias (Sarica, Cerasa, & Quattrone, 2017). Although RFs are time consuming, they are proven to perform better on high-dimensional data (Menze et al., 2009) (X. Chen, Wang, & Zhang, 2011).

5.1.4 Decision Tree Classifier (DTC)

The DTC was introduced in 1977 by Swain and Hauska (1977). A decision tree is a supervised learning algorithm that is used for classification and regression tasks. The best attribute is chosen and placed at the root of the tree. Next, training is divided into subsets based on dataset feature values. Each internal node represents a feature criterion, and each leaf node is a class label. The DTC uses the entire training data to build the model. It is easy to interpret and explain because it is based on if–then rules. Major limitations of the DTC are a long training time, overfitting, large-class problems, and high-dimensional data (Safavian & Landgrebe, 1991).

5.1.5 Perceptron (PN)

Perceptron was introduced by Frank Rosenblatt in 1957 (Rosenblatt, 1958). A perceptron represents a brain neuron cell. A perceptron has an input, weight, weighted sum, and the activation function. Based on input data and weights, the activation function is fired and results in an output. Perceptron is used for binary classification and multi-classification and is easy to implement and train. However, perceptron performs best when classes are separable.

5.1.6 Artificial Neural Networks (ANN)

ANNs are made of many neurons. The neuron network represents human brain neuron cells and how they are connected (Hinton, 1992; Hertz, 2018). ANNs have multiple neurons that are interconnected (Sarle, 1994). A simple neural network is made of an input layer, an output layer, and a hidden layer. A more complex neural network can have multiple hidden layers with a huge number of neurons in each layer.

ANNs are very powerful algorithms; however, one big limitation are their large data requirement. To perform better and produce good results, ANNs require huge amounts of data (Tu, 1996). Moreover, training time increases with the complexity of the network and data. Another disadvantage is the large number of hyperparameters when tuning the ANN.

5.1.7 Deep Learning (DL)

DL has emerged in 2006 as a more complex and deeply structured form of ANNs (Goodfellow, Bengio, Courville, & Bengio, 2016). The difference and complexity are due to the different number of neurons and the number of layers. It also selects different features to fire selected neurons in each layer. DL has different variations (Ravì et al., 2017), such as deep belief networks, CNNs, deep auto-encoder, deep Boltzman machine, and recurrent neural network. DL has been successful in many areas, including computer vision, text, and speech. However, there are a few disadvantages associated with the use of DL (Zohuri & Moghaddam, 2020). A major limitation is the huge number of training samples required to achieve a well-built model. This limitation becomes more complex if the class samples are intertwined and have low separability. Another issue is that DL is a black box algorithm, which cannot explain its reasoning. Furthermore, in feed-forward networks, errors can be exponentially high because the network cannot retrace its steps back and correct its current built.

5.1.8 Bagging Classifier (BC)

The bagging classifier was introduced to enhance tree-based algorithms (Breiman, 1996). Samples are withdrawn with replacement from the dataset. The classifier builds several models and predicts the output using majority or voting from all models in consideration. Bagging is the first form of an RF. Bagging uses all features when building its models as opposed to RFs, which uses a subset of features to build multiple decision trees and aggregate them. Bagging reduces variance, but it may overlook high-and low-performing models if prediction is based on voting or aggregation methods.

5.1.9 Logistic Regression (LR)

LR is a supervised machine learning algorithm (King & Zeng, 2001; Kleinbaum, Dietz, Gail, Klein, & Klein, 2002). It is based on probability by limiting its cost function to 0 or 1. LR is a very cost-efficient algorithm. However, it performs badly when dealing with a high-dimensional dataset. Another data-dependent feature is its performance with linear data compared to non-linear data. High data separability tends to be highly advantageous for LR performance.

5.1.10 Adaptive Boosting (Adaboost)

Adaboost stands for adaptive boosting. It is a well-designed boosting algorithm that is based on building several weak classifiers to create a single strong classifier. First introduced by Freund (1995), the concept relies on re-weighting training instances and determining their probability for choosing them as part of the training set. One advantage of Adaboost is its ability to avoid overfitting the data. However, it requires very good quality data without any noise or outliers. Adaboost learns progressively, and so noise and outliers can cause it to perform badly.

5.1.11 Passive Aggressive (PA)

PA is a highly complex classifier and is not as popular as other classifiers. It was introduced by Crammer, Dekel, Keshet, Shalev-Shwartz, and Singer (2006). Its function is indicated by its name: if the model prediction is correct, then no changes are applied to the model, and hence it is passive. If the model prediction is incorrect, then it forces a change in the model to alter its next prediction; therefore, it is aggressive. Due to its incremental behaviour, the advantage of the PA classifier is its small training memory. Unfortunately, this is also a disadvantage because it does not capture the whole idea or flow of the data. Incremental changes can be caused by the order in which the data are presented.

5.1.12 Linear Discriminant Analysis (LDA)

LDA is a technique usually used to reduce dimensionality (Balakrishnama & Ganapathiraju, 1998). However, data must be normally distributed before LDA can be used as a classifier. LDA uses the Bayes theorem (Joyce, 2003) to estimate the probability of an input belonging to a class. It is focused on maximising the means between the classes and minimising the variance within a class. LDA works very well with large and highly separable datasets. However, high-dimensional datasets and overlapping classes hinder its efficiency.

5.1.13 Quadratic Discriminant Analysis (QDA)

QDA is another form of LDA (Mika, Ratsch, Weston, Scholkopf, & Mullers, 1999). It uses a quadratic decision boundary to separate the classes. Like its relative, it has the same limitations when dealing with high-dimensional datasets.

5.1.14 Ridge Classifier (RC)

Ridge regression was first introduced by Arthur Hoerl and Robert Kennard. It was an alternative to the instability of a linear regression when dealing with large datasets (Drucker et al., 1997). It performs feature-weight updates, and the loss function has an additional squared term. It drives down the overall size of the weight values during optimisation and reduces overfitting. Its advantage is not overfitting the model. However, this can be a disadvantage because it trades variance for bias and shrinks coefficients to zero.

5.1.15 Stochastic Gradient Descent (SGD)

SGD randomly picks one data point from the whole data set in each iteration to reduce the amount of computations enormously (Bottou, 2010). It works very well with large datasets, and it is computationally fast. Unfortunately, it requires extensive hyperparameter tuning, and it is very sensitive to feature scaling.

5.1.16 Gradient Boosting (GB)

GB is a type of machine learning boosting (J. H. Friedman, 2002). It relies on the perception that the best possible next model minimises the loss error when combined with previous models. Gradient descent similar to Adaboost relies on empowering the weak classifiers to create a strong classifier. However, as trees are added, existing trees are not removed. It is robust to missing data, requires no scaling, has highly correlated features, and removes irrelevant features in much the same way as RF. It naturally assigns feature importance scores. Its drawbacks include the effect of outliers, which can cause overfitting, and due to the large number of trees it can be computationally expensive.

5.1.17 Hyperparameter Tuning

Machine learning algorithms have several hyperparameters which require tuning for the algorithm to perform and produce its best possible score (Idris, 2016). For example, κNN a hyperparameter is κ value, RF hyperparameter is the number of trees, SVC hyperparameter is the C value. Tuning ML algorithm hyperparameters can be done through three different methods: manually, random search, and grid search. Manually is when the codes -programmer- tunes and changes the value of a parameter manually and running the code with every tune (trial-and-error). Random search and grid search are part of the Python library and are used to automatically tune the parameters of the ML algorithm. Grid search, as the name entails, creates a grid of all the possible combinations of all the parameters and every grid combination model is built. Random search, similar to grid research, but only randomly selecting several combinations to build the ML models. Grid search may require more time and processing power, but it explores all possibilities of parameters and their values while random and manual searches may miss some good combinations of parameters and their values because they are random choices.

Hyperparameter tuning can be fixed on a scoring metric, such as: accuracy, macro recall, $F1$, and precision. Throughout this thesis any mention to parameters is reference to hyperparameters.

5.2 Acoustic Feature Extraction

OpenSMILE (Eyben, Weninger, Gross, & Schuller, 2013) is an extensive open-source toolkit for feature extraction from audio. It is targeted toward audio analysis in speech and music applications. This includes speech recognition, emotion recognition, and speaker identification. In this research, OpenSMILE is the preferred extraction software. This is due to the large number of features that can be extracted from audio data. In addition, OpenSMILE has been developed for commercial use and introduced to Python as a library instead of as standalone software. This has made it easier to extract features from within the Python environment.

Other extraction software is focused on acoustic features, such as energy, pitch, and speech rate, or spectral features, such as MFCC and cepstral features. OpenSMILE stands out by including LLD, which have influenced the research in the area of speech.

The configuration file used is compare2016.con, which was used in the Interspeech challenge in 2016 (Schuller et al., 2016). Weninger, Eyben, Schuller, Mortillaro, and Scherer (2013b) provide a thorough description of this configuration file.

Table 5.1: *The ComParE acoustic feature set: 65 provided LLD ((Weninger et al., 2013a) - Creative Commons Attribution License 3.0).*

4 Energy Related LLD	**Group**
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
RMS Energy, Zero-Crossing Rate	Prosodic
55 Spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
MFCC 1–14	Cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral roll-off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Psychoacoustic sharpness, harmonicity	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 Voicing Related LLD	**Group**
$F0$ (SHS & Viterbi smoothing)	Prosodic
Prob. of voicing	Voice quality
log. HNR, Jitter (local & delta), Shimmer (local)	Voice quality

5.3 Feature Reduction Techniques

Feature reduction techniques are designed to select a subset of the original features while maintaining useful and necessary information to separate classes. There are many feature reduction techniques, five of which have been chosen for this experiment: principle component analysis, recursive feature elimination, least absolute shrinkage and selection operator, analysis of variance, and RF. Brief descriptions of these techniques are provided in the following sections.

5.3.1 Principle Component Analysis (PCA)

PCA is a dimensionality-reduction technique used for reducing dimensionality while maintaining most of the important information in the downsized feature set (Wold, Esbensen, & Geladi, 1987).

Table 5.2: *The ComParE acoustic feature set: functionals applied to LLD contours, $*$ Arithmetic mean of LLD / positive Δ LLD. $**$ Not applied to voicing related LLD except* F0. *$***$ Only applied to* F0. *((Weninger et al., 2013a) - Creative Commons Attribution License 3.0).*

Functionals applied to LLD/Δ LLD	**Group**
Quartiles 1–3, 3 inter-quartile ranges	Percentiles
1% percentile (≈ min), 99% pctl. (≈ max)	Percentiles
Percentile range 1 %–99%	Percentiles
Position of min / max, range (max – min)	Temporal
Arithmetic mean$*$, root quadratic mean moments	Moments
Contour centroid, flatness	Temporal
Standard deviation, skewness, kurtosis moments	Moments
Relative duration LLD is above 25 / 50 / 75 / 90% range	Temporal
Relative duration LLD is rising	Temporal
Relative duration LLD has positive curvature	Temporal
Gain of linear prediction (LP), LP Coeff. 1–5	Modulation
Mean, max, min, SD of segment length $**$	Temporal
Functionals applied to LLD only	**Group**
Mean value of peaks	Peaks
Mean value of peaks – arithmetic mean	Peaks
Mean / SD of inter peak distances	Peaks
Amplitude mean of peaks, of minima	Peaks
Amplitude range of peaks	Peaks
Mean / SD of rising / falling slopes	Peaks
Linear regression slope, offset, quadratic error	Regression
Quadratic regression a, b, offset, quadratic error	Regression
Percentage of non-zero frames $***$	Temporal

5.3.2 Recursive Feature Elimination (RFE)

RFE involves the backward selection of features (Granitto, Furlanello, Biasioli, & Gasperi, 2006). It performs a greedy search to find the best subset. Then, it builds the model on the whole feature set, and the least important features are removed. The model is rebuilt again using the remaining feature set, and so on. There are two parameters that can be tuned: the number of features and the model selection algorithm (Guyon, Weston, Barnhill, & Vapnik, 2002).

5.3.3 Analysis of Variance (ANOVA)

This is a statistical technique that can be used for feature reduction (Girden, 1992). ANOVA calculates the variance of the feature set, and any features that are independent of the target classes are removed from the dataset. The F-value measures the linear dependency between the feature variable and the target. However, the F-value may underestimate the relation between a feature and the target if the relationship is non-linear.

5.3.4 Least Absolute Shrinkage and Selection Operator (LASSO)

A feature technique proposed by Tibshirani in 1996 (Bühlmann & van de Geer, 2011), LASSO is based on adding a penalty to the model's parameter to reduce overfitting. This is known as regularisation. LASSO or the $\ell 1$ regularizer is applied to the coefficient. Shrinking the coefficient to zero removes that feature from the model.

5.3.5 Random Forest (RF)

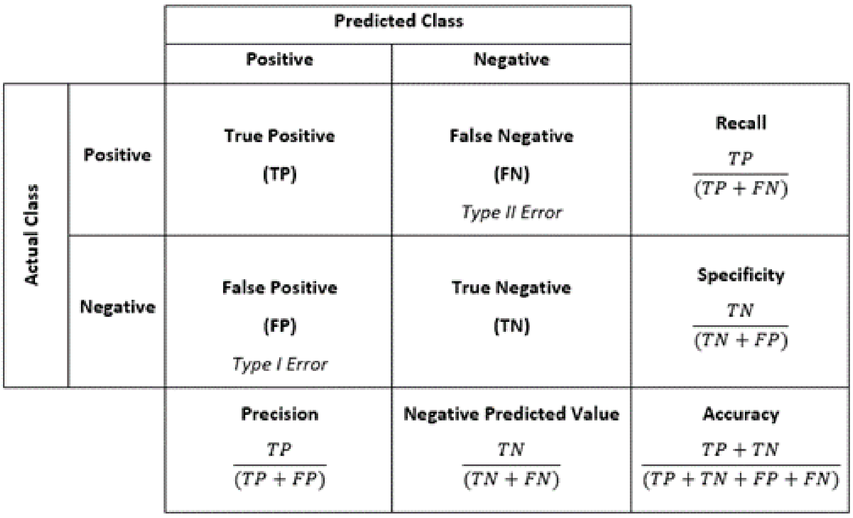
RF is a popular machine learning technique that can be applied to a dataset for feature reduction. RFs provide a feature importance attribute, which assists in choosing the best or important features.

5.4 Evaluation Measures

In binary or multi-class classification, there are several possible evaluation measures. These measures depend on how well the data are balanced. For example, if the data are equally balanced between classes, then it is possible to have an equal probability of appearing in the training sample and testing sample. Therefore, accuracy would be a suitable measure of classifier performance. In contrast, if the dataset is not balanced, then accuracy may be a misleading measure of the classifier's true performance. In this section, several evaluation metrics are described.

Important terms associated with the confusion matrix, shown in Figure 5.1 are explained as follows:

- ***TP***: True Positives: class instances that are correctly predicted as true.
- ***TN***: True Negatives: class instances that are true but predicted as false.
- ***FP***: False Positives: class instances that are correctly predicted as false.
- ***FN***: False Negatives: class instances that are false but predicted as true.



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) *Type II Error*	Recall $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) *Type I Error*	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predicted Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 5.1: *Confusion matrix in machine learning.*

5.4.1 Accuracy

The most common measure is accuracy. Accuracy considers only the correctly predicted labels. It is the ratio of correct predictions to the total number of predictions. Accuracy does not reflect incorrect predictions. Moreover, it performs poorly when dealing with imbalanced datasets, such as if a dataset has 9:1 ratio of two classes. A classifier predicting 100 instances will always predict the majority class, and eventually the classifier will achieve an accuracy of 90%. This is an incorrect measure because it does not have enough data to train on the minority class.

Due to its inappropriateness for imbalanced data, accuracy, although calculated, will not be considered as a performance measure for classifiers. Accuracy is determined as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

5.4.2 Recall

This is the measure that calculates the number of positively predicted classes out of positive class instances and true and false positives. Recall is a better measure of classifier performance than accuracy because it calculates the true number of class instances that are classified correctly from all positive instances. In the event of imbalanced datasets, recall is a better measure of the true classifier's performance. Recall is calculated as follows:

$$recall = \frac{TP}{(TP + FN)}$$

5.4.3 Precision

Precision is a measure that reveals how many of the positively predicted classes are predicted correctly. As the name indicates, 'How precise is your prediction?'

Used together, precision and recall are more powerful than relying on either one separately or relying on accuracy alone. Both measure how well a classifier can predict a class label and how precise it is in classifying the positive class label. Precision is calculated as follows:

$$precision = \frac{TP}{(TP + FP)}$$

5.4.4 *F*1

This measure combines both recall and precision. However, it does not calculate the arithmetic mean. In contrast, it is the harmonic mean of both measures. The harmonic mean refers to when one measure is very low and the other is very high, such as high recall and low precision. It will lean toward the smaller number to indicate the actual performance of the classifier. $F1$ is calculated as follows:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

5.4.5 Area Under the Curve (AUC)

A very important and common measure in machine learning is AUC. It captures the classifier's ability to distinguish between classes. An AUC score of 50% indicates that the classifier cannot separate the classes. Meanwhile, a high score is proof of its ability to separate and distinguish between the classes.

5.5 Statistical Significance

5.5.1 Mathews's Correlation Coefficient (MCC)

MCC is a more reliable statistical measure that reflects the classifier performance. It is a reflection of the full confusion matrix. MCC has a value between -1 and 1. A score close to +1 indicates a good classifier performance. MCC takes into account the four possibilities of a confusion matrix: *TP*, *FP*, *TN*, and *FN* (Chicco & Jurman, 2020). MCC is calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5.5.2 ρ-Value

Another important measure of classifier performance is its significance. Significance indicates whether the classifier performs well and is not producing random results by chance. The ρ-value is calculated from the confusion matrix and the no information rate (NRI). NRI is the classifier's performance based on majority class. The ρ-value is calculated using a one-sided binomial test.

5.6 Classifiers' Baseline

To evaluate ML classifiers, their performance is tested against a naive classifier. The naive classifier is the baseline which acts as a benchmark for classifiers to perform better than a naive classifier. The SKLearn Python library provides a dummy classifier (Idris, 2016). Throughout this research, dummy classifier was used to create the baseline for all experiments.

Chapter Summary

This chapter described the setting of the experiments that were conducted throughout the research. It also described several machine learning algorithms, OpenSMILE, the acoustic features extraction tool, feature reduction techniques, and the evaluation measures.

It was apparent that $F1$ score and accuracy can be misleading because they do not take into account the full confusion matrix (Akosa, 2017). High recall and high precision can indicate that the model performs well. Therefore, this study has adopted recall, precision, and AUC. For significance, it used MCC (G. Liu et al., 2013; J. Yang, Roy, & Zhang, 2013). All performance measures used throughout this research are macro measure that considers all classes.

Chapter 6

Experiments with the Speaker Personality Corpus

> *All life is an experiment. The more experiments you make the better.*
>
> Ralph Waldo Emerson

Personality recognition requires large datasets or repositories (Finnerty et al., 2016; Mohammadi & Vinciarelli, 2012). The largest available one is the SPC (Mohammadi et al., 2010) (see Chapter 4, section 4.1). Numerous studies have used this corpus.

The SPC is a speech corpus that was designed to study automatic personality perception (APP), which is completely different from automatic personality recognition (APR). The former is focused on how personality is perceived by other people, while the latter is the recognition of the speaker's personality.

Mohammadi et al. (2010) experimented with this corpus. Praat (Boersma & Van Heuven, 2001) was used to extract acoustic features. The experiments were conducted using three assessors and SVM with the radial bias function kernel and cross-validation with $k = 15$. The results indicated that high inter-rater agreement was correlated with a high personality score.

6.1 Experiment

The SPC is the largest corpus available that measures perception. However, it was incorrectly used for APR in several conferences in the challenge track. The evaluation metric used for these challenges was unweighted average recall (UAR). There was no measure of significance, and there was no emphasis on AUC.

This section describes how the experiments were executed. The software used for feature extraction was OpenSMILE (Eyben et al., 2013). OpenSMILE is an extensive open-source toolkit for feature extraction from audio. OpenSMILE extracted 6737 features from the SPC using the comapre2016 configuration setting.

The five most-common machine learning algorithms were selected for this experiment: κNN, LR, SVM, RF, and DL. Data were split into 70% training and 30% testing. A model baseline was created before the experiments to serve as the measurement for classifier performance. This was a binary-class problem, and thus the recall and AUC baselines were 50% for both measures.

The machine learning algorithms were trained and tested on the SPC. Prior research preferred accuracy as an evaluation measure and cross-validation with $k = 15$. In recent research, most results have not been significant. This is clearly shown in the initial replication results. The experiments in this study included the error rate. The error rate is the number of incorrect predictions compared to all predictions made. The experiments were divided into five experiments. Experiment 1 involved applying machine learning algorithms without any parameter tuning, and the training was fit on accuracy. Experiment 2 included machine learning algorithms with parameter tuning for best accuracy performance. Experiment 3 used machine learning algorithms without any parameter tuning, and the training was fit on recall. Experiment 4 applied machine learning algorithms with feature reduction, and the training was fit on recall. Experiment 5 utilised machine learning algorithms with parameter tuning for best recall performance.

6.1.1 Experiment 1: Accuracy Models

Experiment 1 explored all possible accuracy fit results without hyperparameter tuning the machine learning parameters. Classifiers were used with default parameter settings. Table 6.1 presents the openness accuracy scores from five different machine learning algorithms. It is clear that SVM performed above the baseline with an accuracy of 63%. However, AUC was below 50%, and this is an indication of the randomness of the results from the classifier. This is further confirmed by the MCC score below zero, which reflects the randomness.

Table 6.1 shows conscientiousness accuracy scores from five different machine learning algorithms. Similar to the openness trait, SVM performed above the baseline with an accuracy of 63%. Moreover, AUC had a score of 50%, which is an indication of the randomness of the results from the classifier. The MCC score was 'None', which means the MCC result was divided by zero. The extraversion accuracy score was highest with logistic regression (55%). Similar to previous traits, AUC and MCC scores are a sign of the classifier's randomness.

The agreeableness trait accuracy score was close to the baseline. DL was the only algorithm to pass the baseline, with an accuracy of 51%. All other classifiers failed to reach the baseline and were deemed unsuccessful. However, it was evident that DL AUC was a sign of the classifier's random results. The neuroticism trait performed the worst of all big five personality traits. It was shown that only κNN and RF scored 50%, which was the baseline score. In addition, the AUC and MCC scores reflected randomness.

Table 6.1: *OCEAN traits scored based on accuracy fit with no hyperparameter tuning applied. The highest accuracy score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	**Training ACC**	**ACC**	**Recall**	**AUC**	**ρ-value**	**Error Rate**	**MCC**
κNN -O	54.22%	59.43%	49.73%	49.72%	0.96	0.963	0
LR -O	58.16%	58.02%	53.34%	53.33%	0.986	0.41	0.06
RF -O	62.62%	61.32%	52.12%	52.11%	0.889	0.38	0.05
DL -O	55.59%	58.49%	54.33%	54.32%	0.980	0.41	0.08
SVM -O	64.49%	**63.21%**	49.49%	49.49%	0.742	0.36	-0.02
κNN -C	60.31%	62.26%	51.80%	51.79%	0.785	0.37	0.04
LR -C	53.77%	58.49%	50.99%	50.98%	0.972	0.41	0.02
RF -C	57.27%	61.32%	53.18%	53.17%	0.859	0.38	0.07
DL -C	53.32%	56.13%	51.27%	51.27%	0.995	0.43	0.02
SVM -C	63.80%	**64.62%**	50.00%	50.00%	0.531	0.35	None
κNN -E	45.80%	52.83%	52.41%	52.41%	0.527	0.47	0.04
LR -E	47.40%	**55.66%**	55.46%	55.46%	0.224	0.44	0.10
RF -E	49.72%	53.77%	53.46%	53.46%	0.418	0.46	0.06
DL -E	48.35%	50.00%	50.48%	50.48%	0.814	0.50	0
SVM -E	47.86%	50.00%	48.88%	48.87%	0.814	0.50	-0.02
κNN -A	50.44%	49.06%	49.06%	49.05%	0.634	0.50	-0.01
LR -A	51.17%	49.53%	49.53%	49.52%	0.581	0.50	0
RF -A	48.52%	45.28%	45.28%	45.28%	0.925	0.54	-0.09
DL -A	53.04%	**51.89%**	51.89%	51.88%	0.315	0.48	0.03
SVM -A	50.53%	48.11%	48.11%	48.11%	0.731	0.51	-0.03
κNN -N	54.67%	50.00%	49.75%	49.75%	0.775	0.50	0
LR -N	48.85%	45.75%	45.61%	45.61%	0.976	0.54	-0.08
RF -N	47.67%	**50.47%**	50.21%	50.20%	0.732	0.49	0
DL -N	50.50%	45.75%	45.61%	45.61%	0.976	0.54	-0.08
SVM -N	45.09%	45.75%	45.34%	45.34%	0.976	0.54	0

Figure 6.1 displays the performance of all classifiers for all personality traits.

Openness and conscientiousness are the only traits that outperformed the remaining traits above the baseline of 50%. The remaining three traits barely exceeded the baseline. Therefore, they might be the hardest to perceive without hyptertuning the best parameters or applying feature reduction techniques.

Detailed figures corresponding to each trait can be found in Appendix A.

6.1.2 Experiment 2: Accuracy Models with Hypertuning

Experiment 2 did not use the naive classifiers. It aimed to further enhance the classifiers' performance through hyperparameter tuning. Further technical details are in Appendix E.

Table 6.2 shows the results of hyperparameter tuning and fitting the model on best accuracy. It displays relatively higher performance with hyperparameter tuned classifiers than with naive classifiers.

However, the extraversion, agreeableness, and neuroticism traits increased slightly when the classifiers' parameters were hyperparameter tuned on the training set to produce the best-fit model with the highest training accuracy.

Figure 6.2 shows the overall performance enhancement for all personality traits, especially openness and conscientiousness. However, extraversion, agreeableness, and neuroticism remain the worst-performing traits despite a slight improvement in their accuracy scores.

Detailed figures corresponding to each trait can be found in Appendix A.

6.1.3 Experiment 3: Recall Models

In the third experimental setting, all models were trained to achieve the best recall fit. There were no hyperparameter tuning or feature reduction techniques applied, and so this serves as a baseline for experiments 4 and 5.

Table 6.3 shows that regardless of the type of classifier, all traits performed around the baseline, with an average accuracy of 56.06%. The AUC and MCC scores clearly indicate classifier randomness when predicting class labels. Meanwhile, agreeableness was the best-performing trait with DL and had the only significant result ($\rho < 0.01$).

Figure 6.3 shows the overall performance of all five personality traits.

Table 6.2: *OCEAN traits scored based on accuracy fit with hyperparameter tuning applied. The highest accuracy score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	ρ- value	Error Rate	MCC
κNN -O	65.67%	65.09%	50.63%	50.62%	0.531	0.34	0.04
LR -O	65.67%	65.09%	50.63%	50.62%	0.531	0.34	0.04
RF -O	64.96%	**65.57%**	51.62%	51.61%	0.474	0.34	0.08
DL -O	59.61%	57.55%	51.41%	51.41%	0.990	0.42	0.02
SVM -O	65.42%	65.09%	50.00%	50.00%	0.531	0.34	None
κNN -C	65.42%	63.68%	49.27%	49.27%	0.642	0.36	-0.07
LR -C	63.31%	**64.62%**	50.30%	50.30%	0.531	0.35	0.02
RF -C	62.82%	64.15%	52.35%	52.35%	0.587	0.35	0.07
DL -C	58.44%	58.96%	50.75%	50.74%	0.962	0.41	0.01
SVM -C	64.48%	64.62%	50.00%	50.00%	0.531	0.35	None
κNN -E	53.09%	**55.19%**	54.96%	54.96%	0.268	0.44	0.09
LR -E	47.68%	53.30%	53.29%	53.28%	0.473	0.46	0.06
RF -E	50.48%	50.47%	49.64%	49.64%	0.775	0.49	0
DL -E	49.15%	**55.19%**	54.86%	54.85%	0.268	0.44	0.09
SVM -E	52.59%	52.83%	50.00%	50.00%	0.527	0.47	None
κNN -A	52.51%	53.77%	53.77%	53.77%	0.151	0.46	0.07
LR -A	53.74%	49.06%	49.06%	49.05%	0.634	0.50	-0.01
RF -A	47.41%	**54.25%**	54.25%	54.24%	0.121	0.45	0.08
DL -A	54.92%	51.42%	51.42%	51.41%	0.365	0.48	0.02
SVM -A	51.18%	51.42%	51.42%	51.41%	0.365	0.48	0.02
κNN -N	57.04%	51.89%	51.65%	51.64%	0.481	0.48	0.03
LR -N	50.76%	42.45%	42.24%	42.23%	0.575	0.57	-0.15
RF -N	50.26%	49.06%	48.50%	48.49%	0.509	0.50	-0.03
DL -N	52.43%	48.11%	47.86%	47.86%	0.51	0.518	-0.04
SVM -N	52.34%	**52.83%**	50.54%	50.53%	0.47	0.471	0.04

Detailed figures corresponding to each trait can be found in Appendix A.

6.1.4 Experiment 4: Recall Models with Feature Reduction

This experiment is divided into three parts to reflect the effect of each feature technique on the classifiers performance. Feature reduction techniques were previously described in Chapter 5. Further technical details are in Appendix E.

6.1.4.1 ANOVA

Table 6.4 shows that the recall score does not improve with the ANOVA feature reduction technique. All traits performed worse than a naive classifier. Figure 6.4 clearly shows that the OCEAN traits perform worse when compared to the baseline. None of the models built produced significant results, and their corresponding MCC values were negligible.

Detailed figures corresponding to each trait can be found in Appendix A.

Table 6.3: *OCEAN traits scored based on UAR fit with no hyperparameter tuning is applied. The highest recall score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	ρ-value	Error Rate	MCC
κNN -O	47.78%	59.43%	49.73%	49.72%	0.963	0.40	0
LR -O	50.73%	58.02%	53.34%	53.33%	0.986	0.41	0.06
RF -O	50.50%	61.32%	52.12%	52.11%	0.889	0.38	0.05
DL -O	52.82%	58.49%	**54.33%**	54.32%	0.980	0.41	0.08
SVM -O	49.51%	63.21%	49.49%	49.00%	0.742	0.36	-0.02
κNN -C	51.39%	62.26%	51.80%	51.79%	0.785	0.37	0.04
LR -C	48.23%	58.49%	50.99%	50.98%	0.972	0.41	0.02
RF -C	51.12%	61.32%	**54.34%**	53.17%	0.859	0.38	0.07
DL -C	50.73%	56.13%	51.27%	51.27%	0.995	0.43	0.02
SVM -C	51.14%	64.62%	50.00%	53.00%	0.353	0.50	None
κNN -E	46.58%	52.83%	52.41%	52.41%	0.527	0.47	0.04
LR -E	47.95%	55.66%	**55.46%**	55.46%	0.224	0.44	0.10
RF -E	51.13%	53.77%	53.46%	53.46%	0.418	0.46	0.06
DL -E	48.84%	50.00%	50.48%	50.48%	0.814	0.50	0
SVM -E	46.20%	50.00%	49.00%	49.00%	0.814	0.50	-0.02
κNN -A	48.50%	56.60%	56.70%	56.69%	0.042	0.43	0.13
LR -A	54.02%	53.77%	53.81%	53.80%	0.185	0.46	0.07
RF -A	48.00%	50.40%	50.41%	50.40%	0.527	0.49	0
DL -A	55.28%	58.96%	**58.97%**	58.97%	0.007	0.41	0.17
SVM -A	51.50%	52.36%	52.32%	52.32%	0.31	0.47	0.04
κNN -N	55.01%	56.13%	55.70%	55.70%	0.26	0.43	0.11
LR -N	49.66%	53.30%	53.43%	53.42%	0.58	0.46	0.06
RF -N	46.77%	57.55%	**57.23%**	57.23%	0.15	0.42	0.14
DL -N	46.07%	52.36%	52.84%	52.83%	0.68	0.47	0.05
SVM -N	48.94%	49.53%	49.42%	49.41%	0.90	0.50	-0.01

6.1.4.2 LASSO

Similar to ANOVA, LASSO as a feature reduction technique failed to produce recall results higher than a naive classifier, as its clear in Table 6.5. Although agreeableness produced a significant result with $\rho < 0.05$, the classifier's MCC value deemed it negligible.

Figure 6.5 shows the OCEAN traits performance compared to the baseline.

Detailed figures corresponding to each trait can be found in Appendix A.

6.1.4.3 Random Forest

Similar to the previously discussed feature techniques, RF performed and produced similar results to LASSO and ANOVA, as shown in Table 6.6. Conscientiousness was the only trait that performed slightly better with RF as a feature reduction technique and with DL as a model. However, the result was not significant, and the classifier produced random results (Figure 6.6). Detailed figures corresponding to

Table 6.4: *OCEAN traits scored based on UAR fit with ANOVA applied for feature reduction. The highest recall score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	ρ- value	Error Rate	MCC
κNN -O	51.61%	50.47%	47.91%	47.90%	0.999	0.49	-0.04
LR -O	67.91%	54.72%	51.58%	51.57%	0.985	0.45	0.03
RF -O	51.98%	61.79%	51.18%	51.17%	0.530	0.38	0.05
DL -O	67.21%	54.25%	**52.61%**	52.61%	0.989	0.45	0.05
SVM -O	66.33%	53.77%	50.58%	51.00%	0.992	0.46	0.01
κNN -C	52.98%	56.13%	51.07%	51.07%	0.999	0.43	0.02
LR -C	68.38%	56.60%	52.62%	52.61%	0.999	0.43	0.05
RF -C	55.08%	65.57%	52.35%	52.34%	0.831	0.34	0.06
DL -C	69.28%	57.55%	**54.92%**	54.91%	0.999	0.42	0.09
SVM -C	67.44%	57.08%	53.00%	53.00%	0.999	0.42	0.06
κNN -E	52.50%	53.30%	53.00%	53.00%	0.635	0.46	0.06
LR -E	70.72%	54.72%	**54.39%**	54.38%	0.473	0.45	0.08
RF -E	59.94%	49.53%	48.96%	48.96%	0.925	0.50	-0.02
DL -E	71.15%	53.77%	53.68%	53.67%	0.582	0.46	0.07
SVM -E	70.43%	55.19%	55.00%	55.00%	0.418	0.44	0.09
κNN -A	47.45%	50.47%	50.53%	50.53%	0.527	0.49	0.01
LR -A	66.36%	50.94%	50.89%	50.88%	0.472	0.49	0.01
RF -A	58.39%	54.25%	**54.17%**	54.17%	0.151	0.45	0.08
DL -A	68.27%	50.94%	50.92%	50.92%	0.472	0.49	0.01
SVM -A	64.48%	50.94%	50.88%	50.87%	0.472	0.49	0.01
κNN -N	53.73%	50.47%	50.44%	50.43%	0.849	0.49	0
LR -N	68.23%	51.42%	51.32%	51.31%	0.775	0.48	0.02
RF -N	51.71%	53.77%	53.29%	53.29%	0.528	0.46	0.06
DL -N	69.22%	51.89%	51.61%	51.61%	0.732	0.48	0.03
SVM -N	66.82%	53.77%	**53.72%**	53.72%	0.528	0.46	0.07

each trait can be found in Appendix A.

6.1.5 Experiment 5: Recall Models with Hypertuning

This section presents the results produced when hyperparameter tuning was applied with a focus on recall for model fitting. The results shown in Table 6.7 indicated no major improvement after hyperparameter tuning the classifier parameters compared to naive classifiers. This is reflected clearly in Figure 6.7. None of the traits, regardless of classifier, surpassed a recall of 55%. Detailed figures corresponding to each trait can be found in Appendix A. Further technical details are in Appendix E.

6.2 Results

Experiment 1 showed that without tuning parameters and fitting the models using cross-validation with $\kappa = 15$, accuracy ranged between 45% and 64%. Moreover, the MCC and ρ-values show that the classifiers were performing randomly. In

Table 6.5: *OCEAN traits scored based on UAR fit with LASSO applied for feature reduction. The highest recall score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	ρ- value	Error Rate	MCC
κNN -O	47.43%	57.55%	**54.34%**	54.33%	0.909	0.42	0.08
LR -O	47.43%	58.49%	49.45%	49.44%	0.855	0.41	-0.01
RF -O	51.42%	61.32%	50.09%	50.08%	0.585	0.38	0
DL -O	50.00%	61.79%	50.00%	50.00%	0.530	0.38	None
SVM -O	50.00%	61.79%	50.00%	50.00%	0.530	0.38	None
κNN -C	51.75%	55.19%	48.37%	48.37%	0.999	0.44	-0.03
LR -C	53.14%	63.21%	48.21%	48.21%	0.953	0.36	-0.05
RF -C	53.76%	67.92%	**53.27%**	53.26%	0.591	0.32	0.10
DL -C	49.94%	68.40%	50.00%	50.00%	0.533	0.31	None
SVM -C	50.00%	68.40%	50.00%	50.00%	0.533	0.31	None
κNN -E	50.55%	50.47%	50.56%	50.55%	0.879	0.49	0.01
LR -E	51.02%	50.94%	48.01%	48.00%	0.849	0.49	-0.05
RF -E	58.34%	55.66%	**54.61%**	54.61%	0.365	0.44	0.09
DL -E	50.00%	54.25%	50.00%	50.00%	0.528	0.45	None
SVM -E	48.62%	49.53%	47.00%	47.00%	0.925	0.50	-0.08
κNN -A	57.28%	53.77%	53.79%	53.78%	0.185	0.46	0.07
LR -A	47.52%	51.89%	52.04%	52.04%	0.365	0.48	0.04
RF -A	54.58%	57.08%	**57.03%**	57.03%	0.031	0.42	0.14
DL -A	50.00%	50.47%	50.00%	50.00%	0.527	0.49	None
SVM -A	49.58%	49.53%	49.38%	49.37%	0.634	0.50	-0.01
κNN -N	53.35%	46.70%	46.93%	46.92%	0.983	0.53	-0.06
LR -N	48.59%	46.70%	45.50%	45.49%	0.983	0.53	-0.09
RF -N	53.20%	46.70%	46.29%	46.28%	0.983	0.53	-0.07
DL -N	50.48%	54.72%	**51.02%**	51.02%	0.418	0.45	0.10
SVM -N	49.77%	53.77%	50.00%	50.00%	0.528	0.46	None

experiment 2, hyperparameter tuning did not have a large effect on accuracy, which remained between 42% and 65%. This might be because accuracy as a misleading evaluation measure due to the imbalanced dataset.

Experiment 3 was based on training and fitting on the best recall score and cross-validation with $\kappa = 10$. Although recall was a good performance measure, the classifiers failed to perform well, and the recall remained between 49% and 58%. Additionally, the MCC and ρ-values indicated the randomness and instability of the classifiers.

Experiment 4 applied machine learning algorithms and performed cross-validation with $\kappa = 10$. The scoring parameter that was used for selecting the best model was macro recall. Three feature reduction techniques were applied. However, the scores ranged between 44% and 57%. Moreover, the MCC and ρ-values indicated that the results were not significant, and the classifiers' performance was random.

Experiment 5 applied machine learning algorithms hyperparameter tuning and cross-validation $\kappa = 10$, and the scoring parameter for selecting the best model was macro recall. The scores ranged from 44% to 54%. In addition, the MCC

Table 6.6: *OCEAN traits scored based on UAR fit with random forest applied for feature reduction. The highest recall score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	$\rho-$value	Error Rate	MCC
κNN -O	51.70%	49.06%	44.64%	44.64%	0.999	0.50	-0.11
LR -O	57.94%	54.25%	50.02%	50.01%	0.989	0.45	0
RF -O	51.78%	61.79%	50.94%	50.94%	0.530	0.38	0.04
DL -O	56.09%	52.83%	**52.17%**	52.17%	0.996	0.47	0.04
SVM -O	56.31%	53.77%	49.64%	50.00%	0.992	0.46	0
κNN -C	51.07%	55.66%	50.73%	50.72%	0.999	0.44	0.01
LR -C	61.77%	58.49%	52.39%	52.39%	0.999	0.41	0.04
RF -C	54.03%	65.57%	52.75%	52.74%	0.831	0.34	0.07
DL -C	60.76%	57.08%	**54.17%**	54.16%	0.999	0.42	0.07
SVM -C	58.50%	58.02%	52.00%	52.00%	0.999	0.41	0.04
κNN -E	51.40%	57.08%	**57.21%**	57.20%	0.224	0.42	0.14
LR -E	64.28%	53.30%	53.00%	53.00%	0.635	0.46	0.06
RF -E	57.51%	48.58%	47.45%	47.44%	0.957	0.51	-0.05
DL -E	64.61%	56.13%	56.10%	56.09%	0.315	0.43	0.12
SVM -E	64.00%	52.83%	53.00%	53.00%	0.685	0.47	0.05
κNN -A	46.80%	50.47%	50.53%	50.53%	0.527	0.49	0.01
LR -A	58.42%	52.83%	52.83%	52.83%	0.268	0.47	0.05
RF -A	53.18%	53.77%	**53.70%**	53.69%	0.185	0.46	0.07
DL -A	57.97%	49.06%	49.03%	49.03%	0.684	0.50	-0.01
SVM -A	59.34%	53.30%	53.29%	53.29%	0.225	0.46	0.06
κNN -N	53.73%	48.11%	47.89%	47.88%	0.957	0.51	-0.04
LR -N	53.73%	52.36%	52.48%	52.47%	0.685	0.47	0.04
RF -N	52.87%	48.11%	47.60%	47.60%	0.957	0.51	0.47
DL -N	54.60%	54.25%	**54.52%**	54.52%	0.473	0.45	0.09
SVM -N	53.52%	53.30%	53.21%	53.21%	0.582	0.46	0.06

and ρ-values indicated that the results were not significant, and the classifiers' performance was random.

The results do not show confidence in the classifiers' ability to perceive (recognise) personality. Moreover, based on these experiments, there was a clear discrepancy between the results achieved and those in the literature based on the SPC.

Therefore, the next step in the experiment was to highlight several issues in the dataset and propose a solution to overcome them.

6.3 Discussion

The SPC was developed by Mohammadi et al. (2010). The aim was to predict whether personality can be perceived from non-verbal cues. However, the corpus was misused in several research papers as a ground truth for personality recognition when it was intended to reflect personality perception from strangers. And the ground truth was based on judges agreement on personality traits (perception). Further

Table 6.7: *OCEAN traits scored based on UAR fit with hyperparameter tuning applied. The highest recall score for each trait is shown in bold. Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).*

Classifier	Training ACC	ACC	Recall	AUC	ρ- value	Error Rate	MCC
κNN -O	50.44%	65.57%	51.30%	51.30%	0.474	0.34	0.08
LR -O	55.58%	51.89%	49.26%	49.25%	0.999	0.48	-0.01
RF -O	49.93%	65.09%	51.25%	51.25%	0.531	0.34	0.06
DL -O	53.64%	51.42%	48.58%	48.58%	0.999	0.48	-0.02
SVM -O	50.01%	57.08%	**52.93%**	53%	0.993	0.42	0.05
κNN -C	54.43%	61.32%	51.37%	51.36%	0.859	0.38	0.03
LR -C	51.96%	58.49%	51.29%	51.28%	0.972	0.41	0.02
RF -C	51.42%	63.21%	**51.92%**	51.92%	0.694	0.36	0.05
DL -C	53.26%	55.19%	50.55%	50.54%	0.998	0.44	0.01
SVM -C	50.00%	64.62%	50.00%	50.00%	0.531	0.35	None
κNN -E	52.92%	55.19%	**54.96%**	54.96%	0.268	0.44	0.09
LR -E	48.64%	51.89%	51.73%	51.73%	0.634	0.48	0.03
RF -E	50.03%	55.19%	54.54%	54.53%	0.268	0.44	0.09
DL -E	49.90%	54.25%	53.86%	53.85%	0.365	0.45	0.07
SVM -E	50.00%	52.83%	50.00%	50.00%	0.527	0.47	None
κNN -A	51.92%	53.30%	**53.30%**	53.30%	0.185	0.46	0.06
LR -A	52.34%	50.00%	50.00%	50.00%	0.527	0.50	0.50
RF -A	46.92%	53.30%	**53.30%**	53.30%	0.185	0.46	0.06
DL -A	52.07%	51.89%	51.89%	51.88%	0.315	0.48	0.03
SVM -A	52.09%	51.42%	51.42%	51.41%	0.365	0.48	0.02
κNN -N	55.25%	51.89%	**51.73%**	51.73%	0.582	0.48	0.03
LR -N	51.95%	46.70%	46.60%	46.60%	0.957	0.53	-0.06
RF -N	48.33%	50.00%	49.22%	49.21%	0.775	0.50	-0.01
DL -N	52.36%	45.28%	44.98%	44.98%	0.983	0.54	-0.10
SVM -N	51.07%	46.70%	46.69%	46.69%	0.957	0.53	-0.06

research on personality, current corpora, and the SPC specifically has revealed several challenges and limitations in computer science regarding the understanding of personality recognition.

This section highlights the major limitations and misconceptions associated with the misuse of SPC.

Several conference challenge tracks have presented the SPC as the dataset to be used for training and testing. However, SPC does not include the ground truth of personality trait or the final score of the personality trait. Users of SPC had to calculate the ground truth from the personality scores provided in the SPC package. The SPC includes the scores of all eleven judges for each trait for each video.

To calculate whether a trait is high or low, several calculation steps must be completed before the final majority voting. An average trait score must be calculated for each judge. Since there are five personality traits, each judge must have five averaged trait scores. Moreover, the steps must be repeated for all judges. For simplicity, the steps are as follows:

1. For Judge j calculate the average trait x score t from all 640 clips (repeated for all traits).
2. Judge j's score for trait x is used to decide if the numeric trait score given by judge j is above the average t; then trait is high, else trait is low (repeated for all clips).
3. The process must be completed for all judges and all traits. In the end, each clip will have eleven scores (either high or low).
4. The class label for each clip is based on the majority of the judges' scores. If at least six judges agree on a class label for a clip, it is classified as a majority.

After calculating the perceived personality score for each instance in the corpus, it is apparent that the challenges inaccurately represented the number of instances for each trait label. This is defined as class imbalance.

The following sections highlight the remaining limitations and challenges of the SPC and personality recognition.

6.3.1 High-Dimensional Data

OpenSMILE (Eyben et al., 2013) extracted 6373 prosodic features from each audio clip. This led to a large number of features for each observation. In traditional datasets, the number of features (parameters/attributes) is more or less equal to the number of observations. However, the age of data continues to grow due to newly emerging technologies. This has had a great impact on research. Today, large amounts of data are captured at relatively low costs (Fan & Li, 2006). This massive data explosion is known as 'high-dimensional data'. This has caused machine learning models to overfit leading to a decline in their performance (Bühlmann & van de Geer, 2011). This extremely large volume of data presents limitations and challenges to current machine learning algorithms, which is referred to as 'the curse of dimensionality' by Richard Belman (Tang, Alelyani, & Liu, 2014). Traditional statistics cannot deal with this massive growth in dimensionality. Therefore, Donoho (2000) suggested that new methods of high-dimensional data analysis are needed to deal with this growth. Moreover, statisticians consider model selection by selecting a subset of possible explanatory variables that will explain the dependent variable (Donoho, 2000).

$$SampleSize(n) < Features(f)$$

The large number of features may affect the classifiers' performance positively or negatively. To understand the effect on performance, this section experiments with classifiers' performance with reduced features. The primary evaluation measure throughout this thesis is macro recall. Any reference to recall means macro recall.

PCA was the first technique to be applied to the SPC. However, due to the large number of features, PCA failed to run successfully. Several platforms, such as *R*, Anaconda, and SPSS, have failed and continue to crash.

RFE is another feature reduction technique. RFE performance is similar to that of PCA. However, regardless of platform, RFE continued to crash after many hours of running the script.

The results presented are compared against recall with the hyperparameter tuned classifier, naïve classifier, RF for feature reduction with the top 2000 features, LASSO and ANOVA feature reduction. Feature reduction did not significantly affect classifier performance. Most classifiers performed worse than with the full feature sets. In fact, all traits were classified better with full feature sets. However, openness, conscientiousness, and extraversion performed slightly better than other traits with feature reduction.

In addition, the MCC values indicated the randomness of the classifiers' performance. The results cannot be used to accept or reject the hypothesis that personality can be recognised from non-verbal cues. The results are summarized in Tables 6.4, 6.5 and 6.6

6.3.2 Judges' Agreement

The corpus included ratings of 11 judges per audio clip. Aggregation between judges can be useful in reducing errors and increasing reliability (Epstein, 1983). However, with the SPC, it was apparent that aggregation reduced the reliability of the ground truth, and the findings were difficult to interpret. Therefore, the focus was on increasing the judges' agreement by selecting a subset of data where the agreement between three judges was higher than the agreement of 11 judges. This step was completed by repeatedly selecting random judges. The agreement measures used were Cronbach's alpha (Bland & Altman, 1997; Tavakol & Dennick, 2011) and inter-class correlation (ICC) (Koo & Li, 2016). Table 6.8 presents the results of the aggregation of 11 judges with the full corpus and the aggregation of three judges with a subset of the corpus. The table shows a slight increase in the judges' agreement for all traits. The low score for agreement may have been caused by individual judge's unreliability and the non-shared meaning system, which limited the consensus among judges. Therefore, in this experiment a subset of the dataset was chosen, which limits the data to 106 samples and three judges.

Table 6.8: *Judges' agreement using Cronbach's alpha and ICC*

Trait	Full Corpus and 11 Judges		Subset Corpus and Three Judges	
	Cronbach's Alpha	ICC	Cronbach's Alpha	ICC
O	0.183	0.175	0.616	0.614
C	0.584	0.546	0.630	0.631
E	0.823	0.786	0.830	0.832
A	0.622	0.596	0.659	0.660
N	0.628	0.563	0.634	0.635
Mean	0.568	0.533	0.674	0.674

6.3.3 Data Imbalance

Class imbalance in a binary classified dataset means that classes are not represented equally. There is a majority class and a minority class (Japkowicz & Stephen, 2002). If observations are clearly separable, class imbalance may not be an issue in binary classification applications (Menardi & Torelli, 2014). However, with growth of dimensionality, data observations have become more complex and intertwined (Hand & Vinciotti, 2003). King and Zeng (2001) showed how logistic regression can be overwhelmed by the majority class and classify new instances to the majority class. Thus, the minority class is ignored.

6.3.3.1 Sampling Techniques

Several techniques have been proposed to deal with data imbalances, either on a data level or a machine learning level (Cieslak & Chawla, 2008). However, due to the various characteristics of each dataset, classifiers become data specific, and certain classifiers work better when dealing with certain datasets. Therefore, data-level solutions we chosen, such as oversampling and undersampling techniques (Estabrooks, Jo, & Japkowicz, 2004).

6.3.3.2 Oversampling

Oversampling duplicates the minority instances to match the number of instances in the majority class. There are drawbacks to this technique (McCarthy, Zabar, & Weiss, 2005). For example, oversampling may increase the chance of overfitting.

6.3.3.3 Undersampling

Undersampling removes instances from the majority class to give it the same number of instances as the minority class. There are drawbacks to this technique as well (McCarthy et al., 2005). A major drawback is that undersampling may remove important instances that affect machine learning.

6.3.3.4 Synthetic Minority Oversampling Technique (SMOTE)

The drawbacks of oversampling and undersampling have led to new, more complex approaches to oversampling and undersampling. SMOTE creates synthetic instances from minority class using κNN (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). New artificial unseen data reduces the overfitting caused by simple oversampling and improves generalisation (Ertekin, Huang, Bottou, & Giles, 2007).

6.3.3.5 TOMEK Links

TOMEK links were first introduced by (Kubat & Matwin, 1997). TOMEK is a sophisticated undersampling technique that removes majority class instances that are borderline, meaning they have characteristics to minority classes (Batista, Prati, & Monard, 2004; Kotsiantis & Pintelas, 2003; Sain & Purnami, 2015; Tantithamthavorn, Hassan, & Matsumoto, 2018). Tomek links can be used for data pre-processing (clean-up).

6.4 Modified Experiment

In this experiment, SMOTE (oversample) was implemented to balance the data, followed by RF for feature reduction. Afterwards, TOMEK (undersample) was applied to increase the separability between the classes and enhance the machine learning process during classification.

For this experiment, four classifiers were chosen: RF, κNN, LR, and SVC. The results of this experiment are summarized in five tables, each representing a personality trait and classification performance of the four classifiers. The dataset was split as 80% training and 20% testing. Because the dataset changed in size, it was difficult to compare the achieved recall with the baseline recall provided by the 'Interspeech2012 Challenge'(Schuller et al., 2012), therefore, the dummy classifier (Idris, 2016) was used to create a new recall baseline, precision baseline, $F1$ baseline, and AUC baseline for the new subset corpus used in this experiment. Higher recall and precision were achieved along with higher $F1$ and AUC for each trait using feature reduction, class-balancing, and RF combined with different balancing techniques. Different parameters were tuned for each classifier per trait. The results include positive MCC results confirming that the model can distinguish well between classes. Further technical details are in Appendix E.

6.5 Results

Tables 6.9, 6.10, 6.11, 6.12, and 6.13 clearly show that all personality traits were better predicted using RF combined with a sampling technique.

In Tables 6.9 and 6.10, RF coupled with TOMEK and SMOTE, respectively, show the best results when compared to the baseline, while the other machine learning algorithms performed worse than or similar to the baseline.

Table 6.11 clearly shows that all machine learning algorithms with the exception of SVC performed better than the baseline, and RF coupled with TOMEK performed the best.

However, in Tables 6.12 and 6.13 only RF with sampling techniques surpass the baseline in agreeableness and neuroticism traits. Other machine learning algorithms barely pass the baseline or fail to even reach it.

Personality perception is based on judges' agreement and sufficient information to determine the traits. In this experiment, it can be concluded that a slight increase in agreement coupled with data balancing and feature reduction yielded better results than the baseline. Although all classifiers can deal with high-dimensional datasets, the κNN, SVC, and LR failed to perform well due to the low separability between instances in the dataset.

Table 6.9: Openness trait classification results

Trait O	**AUC**	**UA Recall (WA Recall)**	**UA Precision (WA Precision)**	**UA F1 (WA F1)**	**MCC**
Algorithm\Baseline	56.47%	66.47% (59.09%)	61.67% (77.12%)	56.86% (62.21%)	
RF + TOMEK	**81.18%**	**81.18% (81.82%)**	**75.24% (85.11%)**	**77.08% (82.77%)**	**0.56**
LR + SMOTE + TOMEK	55.29%	55.29% (63.64%)	54.29% (68.31%)	54.17% (65.53%)	0.09
kNN + SMOTE	55.88%	55.88% (31.82%)	62.50% (82.95%)	30.53% (25.36%)	0.17
SVC + SMOTE + TOMEK	55.29%	55.29% (63.64%)	54.29% (68.31%)	54.17% (65.53%)	0.09

Table 6.10: Conscientiousness trait classification results

Trait C	**AUC**	**UA Recall (WA Recall)**	**UA Precision (WA Precision)**	**UA F1 (WA F1)**	**MCC**
Algorithm\Baseline	56.47%	66.47% (59.09%)	61.67% (77.12%)	56.86% (62.21%)	
RF + SMOTE	**82.35%**	**82.35% (72.73%)**	**72.73% (87.60%)**	**70.54% (74.92%)**	**0.54**
LR + SMOTE	49.41%	49.41% (54.55%)	49.57% (64.49%)	47.62% (58.01%)	-0.01
kNN + TOMEK	64.71%	64.71% (45.45%)	64.71% (83.96%)	45.45% (45.45%)	0.29
SVC + SMOTE	46.47%	46.47% (50.00%)	47.50% (62.50%)	44.37% (54.02%)	-0.05

Table 6.11: Extraversion trait classification results

Trait E	**AUC**	**UA Recall (WA Recall)**	**UA Precision (WA Precision)**	**UA F1 (WA F1)**	**MCC**
Algorithm\Baseline	73.50%	73.50% (72.73%)	72.73% (74.38%)	72.50% (72.95%)	
RF + TOMEK	**79.06%**	**79.06% (77.27%)**	**78.33% (80.45%)**	**77.23% (77.41%)**	**0.57**
LR + TOMEK	75.64%	75.64% (77.27%)	76.79% (77.11%)	76.03% (77.03%)	0.52
kNN + TOMEK	77.35%	77.35% (77.27%)	76.67% (77.88%)	76.84% (77.42%)	0.54
SVC + TOMEK	71.79%	71.79% (72.73%)	71.79% (72.73%)	71.79% (72.73%)	0.43

Table 6.12: Agreeableness trait classification results

Trait A	**AUC**	**UA Recall (WA Recall)**	**UA Precision (WA Precision)**	**UA F1 (WA F1)**	**MCC**
Algorithm\Baseline	50.00%	50.00% (50.00%)	48.21% (58.77%)	47.28% (52.72%)	
RF + TOMEK	**63.54%**	**63.54% (77.27%)**	**72.81% (75.60%)**	**65.08% (74.46%)**	**0.35**
LR + TOMEK	45.83%	45.83% (59.09%)	45.29% (56.79%)	45.45% (57.85%)	-0.08
kNN + SMOTE	47.92%	47.92% (31.82%)	46.49% (55.66%)	30.53% (26.22%)	-0.05
SVC + SMOTE	50.00%	50.00% (72.73%)	36.36% (52.89%)	42.11% (61.24%)	0

Table 6.13: Neuroticism trait classification results

Trait N	**AUC**	**UA Recall (WA Recall)**	**UA Precision (WA Precision)**	**UA F1 (WA F1)**	**MCC**
Algorithm\Baseline	50.00%	50.00% (45.45%)	43.18% (74.59%)	37.14% (53.77%)	
RF + SMOTE + TOMEK	**58.77%**	**58.77% (77.27%)**	**56.94% (80.18%)**	**57.53% (78.59%)**	**0.15**
LR + SMOTE	34.21%	34.21% (59.09%)	40.63% (70.17%)	37.14% (64.16%)	-0.24
kNN + SMOTE	45.61%	45.61% (54.55%)	47.86% (74.59%)	42.71% (61.65%)	-0.06
SVC + SMOTE	39.47%	39.47% (68.18%)	41.67% (71.97%)	40.54% (70.02%)	-0.18

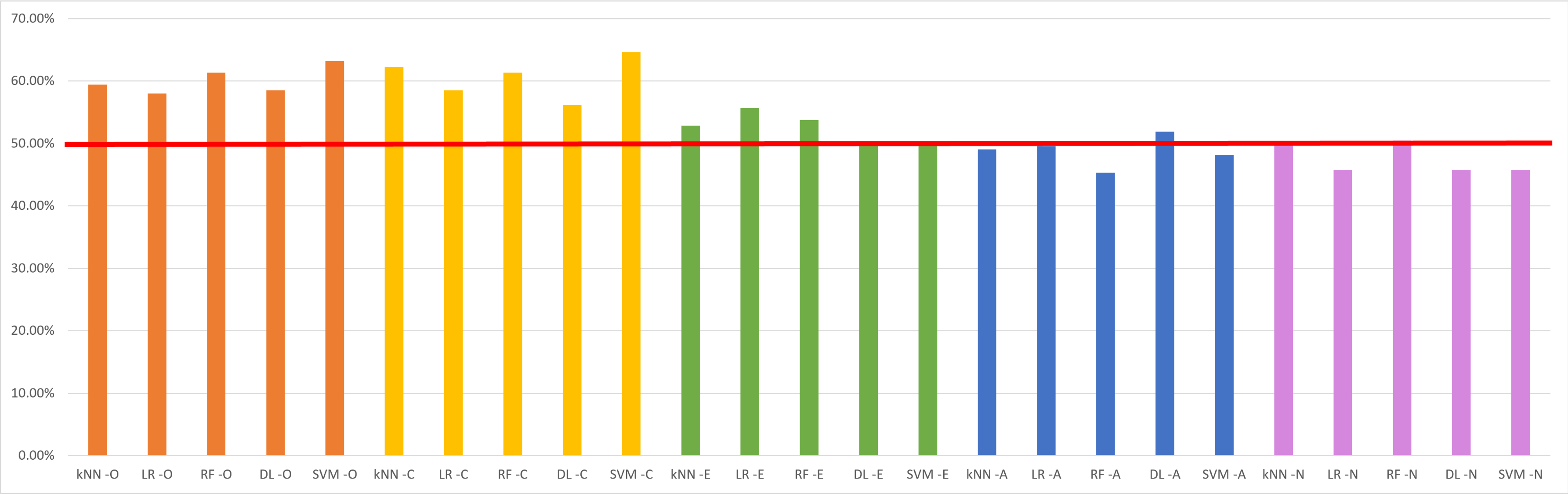


Figure 6.1: *OCEAN traits accuracy measure with no hyperparameter tuning applied to the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

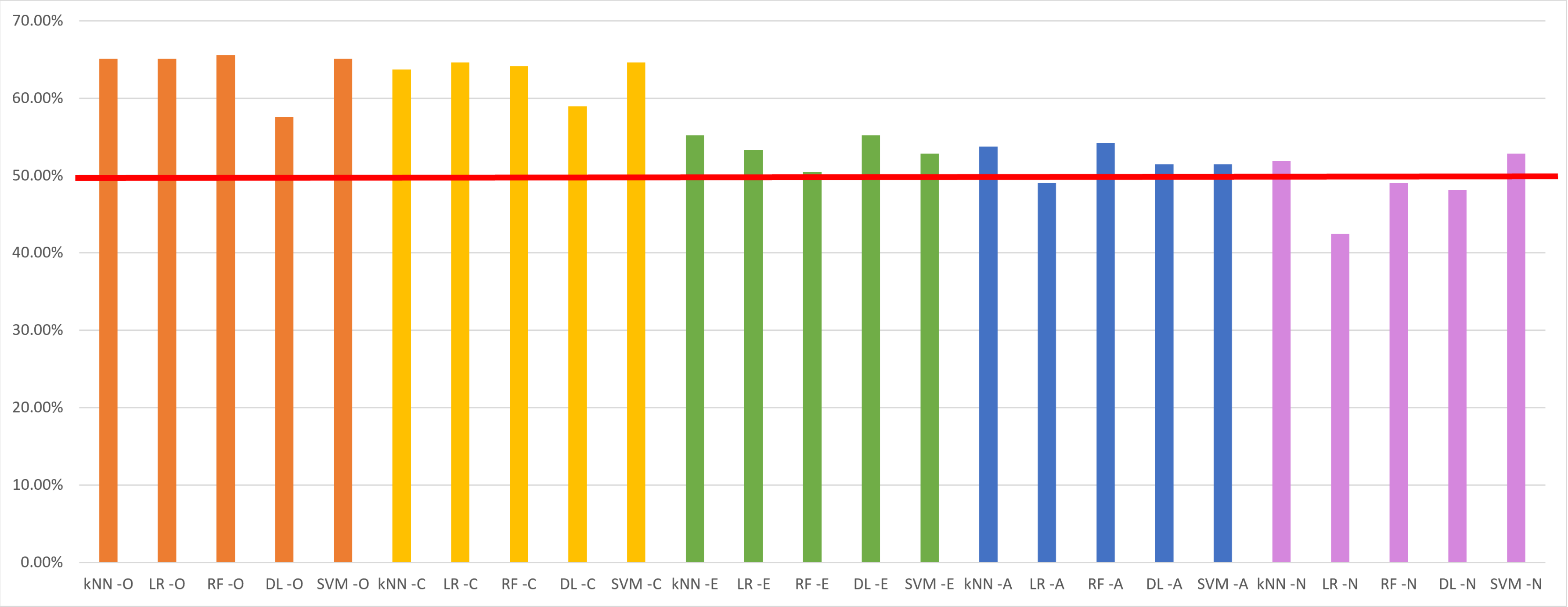


Figure 6.2: *OCEAN traits accuracy measure with hyperparameter tuning applied to select best parameters for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

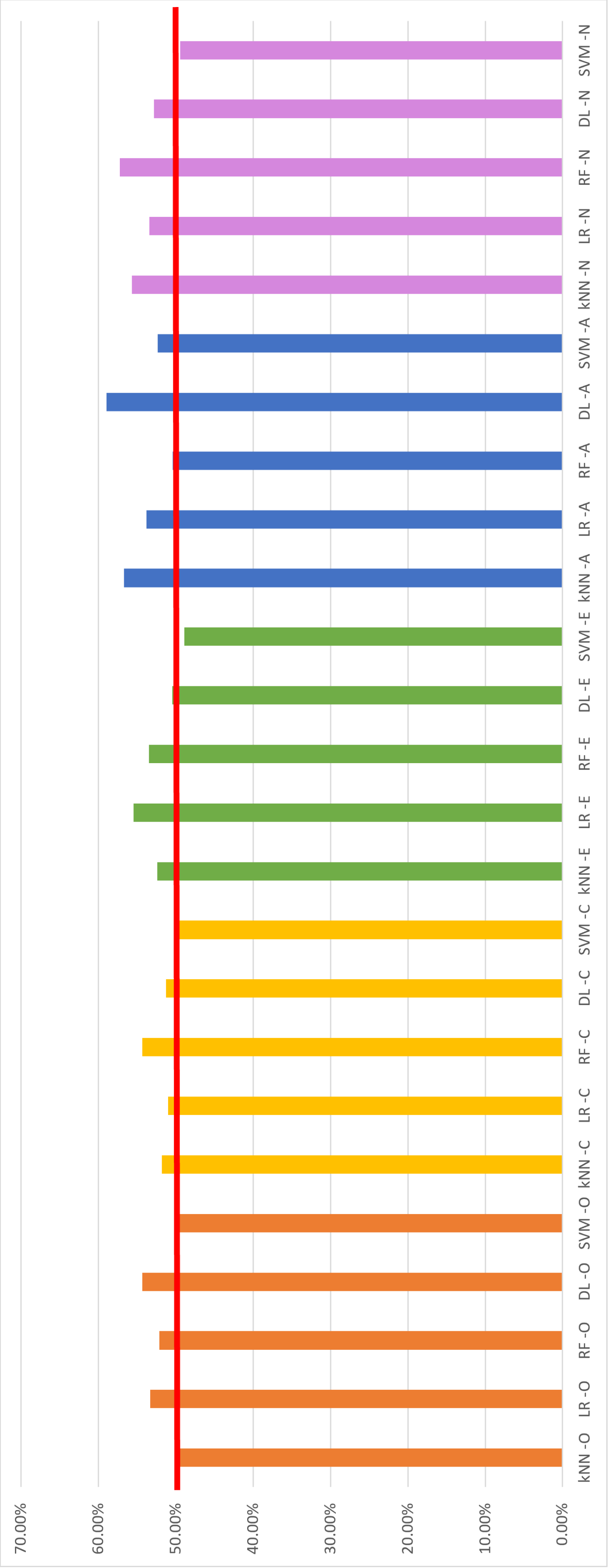


Figure 6.3: *OCEAN traits recall measure with no hyperparameter tuning applied for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

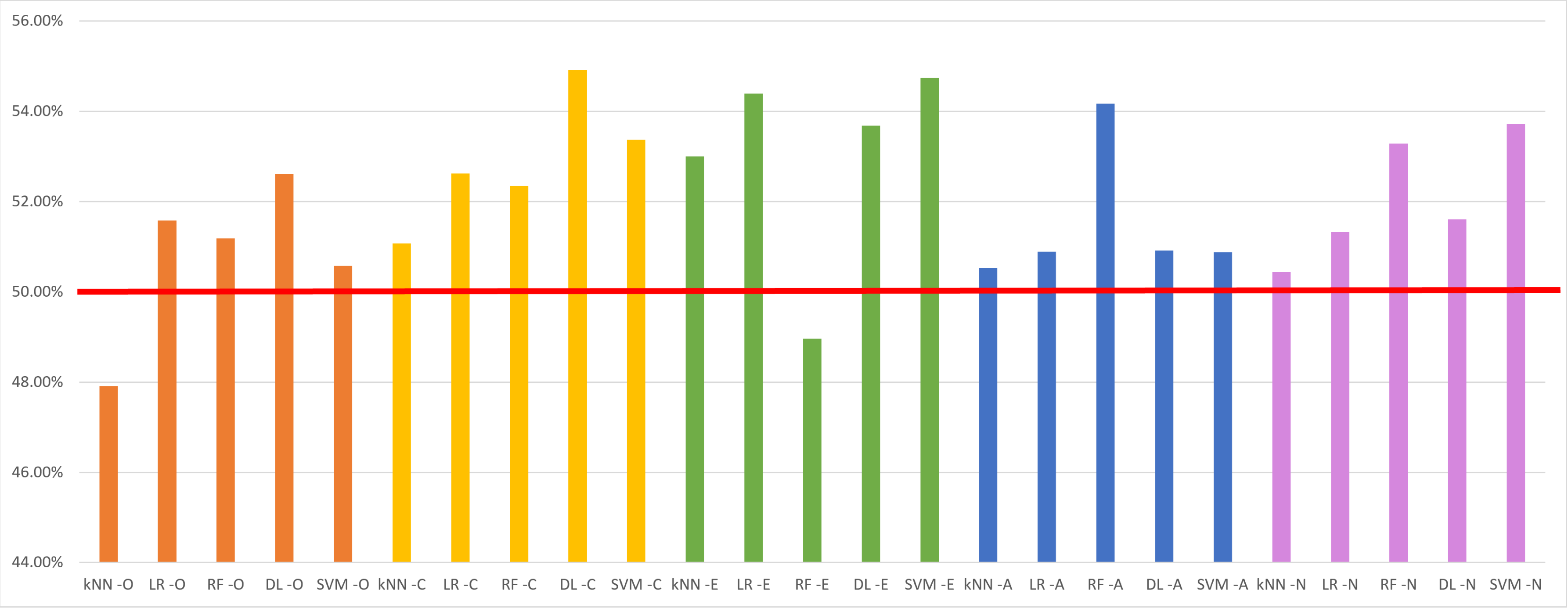


Figure 6.4: *OCEAN traits recall measure with no hyperparameter tuning applied for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

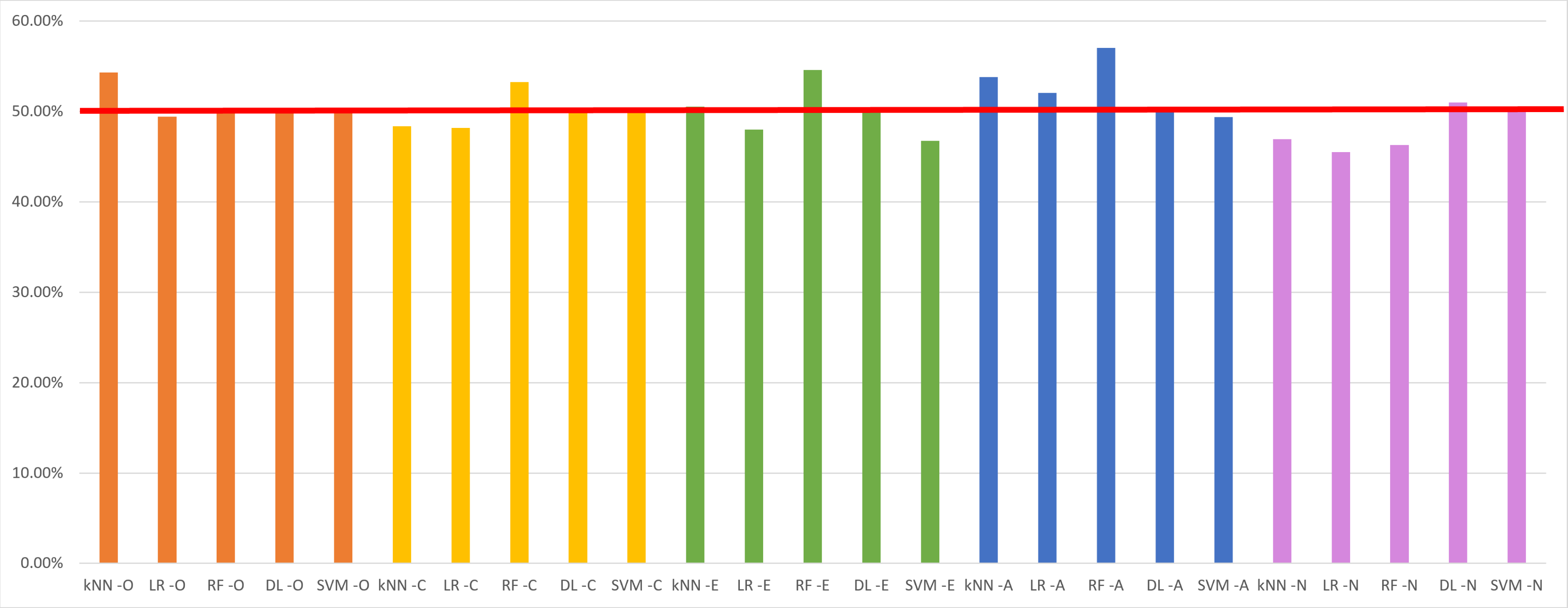


Figure 6.5: *OCEAN traits recall measure with no hyperparameter tuning applied for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

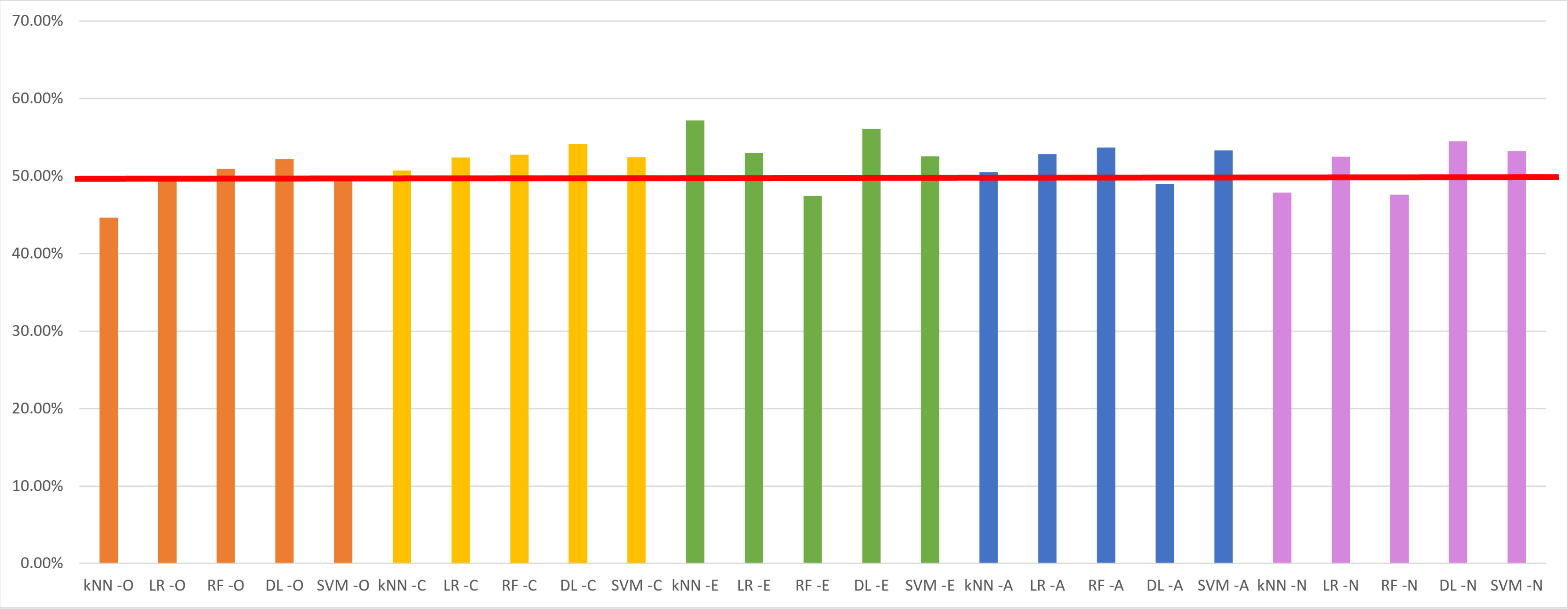


Figure 6.6: *OCEAN traits recall measure with no hyperparameter tuning applied for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

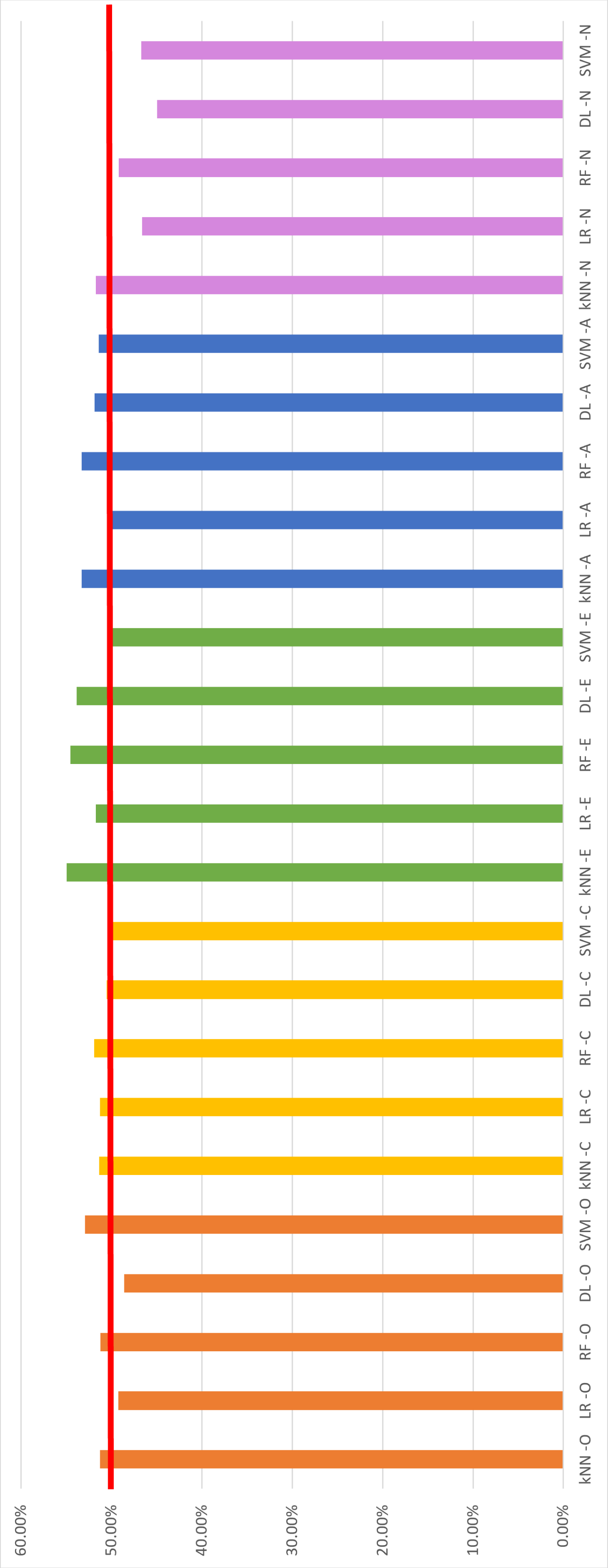


Figure 6.7: *OCEAN traits recall measure and with hyperparameter tuning applied for each of the machine learning algorithms. Openness (orange), Conscientiousness (yellow), Extraversion (green), Agreeableness (blue), Neuroticism (purple)*

Chapter Summary

It can be concluded from this chapter that there were major issues after investigating and experimenting with the SPC. It was found that SPC was misrepresented as a personality recognition corpus when in fact it was a personality perception corpus, which was based on the agreement of zero-acquainted judges. After thoroughly experimenting with the SPC, a discrepancy emerged between the results achieved and those published in the literature.

Based on the last experimental study presented, personality recognition from non-verbal cues is possible. However, the focus should shift to how to achieve higher agreement between judges and to understand how judges can agree on a personality classification based on minimal shared information. Personality recognition in HCI is still new, and further research is required.

There were several limitations that were overcome using different techniques. The high-dimensionality issue was resolved using RF as a feature reduction technique. To overcome class imbalance issue, sampling techniques were used. Lastly, regarding the low consensus between judges, it was increased by selecting a subset of the dataset that included three judges to maintain high agreement.

However, a major issue that has been overlooked when building datasets for personality recognition is the ground truth label of a personality trait. Personality ground truth cannot be accepted from judges who have zero acquaintance with the person being judged.

This is a major issue in current datasets, which are based on acting (demonstrating) a personality trait or being judged by strangers with zero acquaintance, which are used for building personality recognition models. The published research can be misleading. SPC is a personality perception-focused dataset. Using it for personality recognition may thus be inappropriate.

Therefore, this chapter can be concluded with the following statements: First, the currently available personality recognition corpora are not compatible with personality recognition, as they do not have a correctly calculated ground truth. Second, the personality corpora are based on the agreement and consensus of judges and not on the personality of what is being judged. Third, personality traits cannot be acted or faked, thus traits require a stimulus to be evoked.

Chapter 7

Accuracy of Personality Judgement

Wisdom begins in wonder.

Socrates

7.1 Overview

In Chapters 1 and 3, this research presented the definition of personality, its history, and its most popular theories. Based on the experiments and research conducted in Chapter 6 of this thesis, it is clear that accuracy of personality judgement —accuracy is a new and ongoing research area in the psychology community. Most papers published in the computer science community do not reflect the study of accuracy in the psychology community. There is a huge gap, as computer scientists are working on HCI- and personality-related research without realising that accuracy is an ongoing struggle in the psychology community.

Most published research has assumed that personality can be recognised if two or more strangers agree on a personality classification for a single person. For example, strangers $s1$, $s2$, and $s3$ recognise that person $x1$ has the extroversion trait. This is based solely on their agreement, regardless of whether person $x1$ is actually an extrovert. There is no consideration of how accurate their classification is. Moreover, dataset instances were labelled based on majority voting and agreement between strangers and not on correct and accurate personality trait classification. The ground truth calculated for each instance in the dataset was also incorrect.

After extensive research in the psychology community, two issues emerged: the gap between the two communities and the incorrect ground truth understanding. Recently, accuracy in personality has resurfaced in psychology research. Funder (1999) explained that, for the past four decades, psychologists have been in fear of accuracy. They either ignore it or try to re-create it out of existence. In personality, accuracy is not what is perceived but what it is. Funder focused his research on how to reach the true personality so it could be compared with others' judgements.

Why do we care about personality? Or, why is accuracy important? Simple everyday decisions are based on the personality perception of others. When hiring someone as a team leader, interviewers are looking for extroverted, cooperative, leader-type personalities. When hiring a teacher, the focus is someone who is friendly, warm, caring, and honest (Blackman, 2002; Blackman & Funder, 2002; Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005; Lievens, De Fruyt, & Van Dam, 2001). Humans are the only source of data for personality (Funder, 1993). The study of accuracy needs to determine under what conditions judgements become more accurate.

Personality attributes can be inferred through behaviour. How humans act or behave in certain situations, what attributes they show, and how they act are linked to personality traits (Wiggins, 1973).

7.2 Background

Personality psychology and social psychology were born at the same time, and they later became two separate research fields. Gordon Allport was one of the founders of both fields (Funder, 1999). Allport, Vernon, and Powers (1933) examined certain patterns in humans that which uniquely identify them, and they referred to these patterns as 'personality traits'. The authors focused on how 'personality traits' were accurately recognised.

Allport's (1937) focus was not only on personality traits but also on how personality was perceived. He studied how traits were incorporated in both verbal and non-verbal behaviour. Allport believed that if he understood the traits, then he could identify the behavioural cues to accurately recognise and identify these traits.

More than fifty years later, researchers interested in personality have become separated from researchers interested in personality perception, also known as social psychology. Although both deal with personality and are two sides of the same coin, researchers in these two fields prefer to avoid each other. Social psychology has become fixated on experiments on human behaviour, providing and studying different stimuli. Meanwhile, personality psychology has become fixated on correlations and personality questionnaires. The former fails to understand the basics and statistics associated with correlations and the relation of effect size to personality measures. In turn, the latter misunderstand the methods of social psychology and how self-reports are flawed (Funder, 1999).

Personality research was criticised heavily by Mischel (2013). In his book, he clearly stated that the existence of personality traits was minimal, and people perceived traits in each other not because they exist but because people are biased toward perceiving traits regardless of evidence (Funder, 1999).

Mischel's (2013) book has caused a decline in personality research. According to Funder (1999) personality has become a less focused area in psychology due to Mischel's statements in his book.

Funder, among others, rebutted Mischel's statements and argued for the importance and existence of personality traits (Swann Jr & Seyle, 2005).

Another issue personality psychology faced was accuracy of personality judgement. Cronbach (1955) published an article critiquing inter-judge agreement and self–other agreement and claimed that it was tainted by stereotype accuracy. Moreover, he clearly stated that what is perceived is not the unique personality of an individual (differential accuracy), but rather the personality of the average person (normative accuracy). Differential accuracy is now known as distinctive accuracy (Vogt & Randall Colvin, 2003). Cronbach's style of writing has intimidated researchers regarding accuracy and steered them away. Although he did not say it was impossible to accurately recognise personality, his was understood in that way.

These two issues have negatively affected the research on personality and accuracy in psychology, nearly bringing it to a halt. However, as Funder successfully connected the two parts of psychology, he worked with many others to restore interest in personality and, more specifically, judgement, and accuracy.

7.3 Related Work on Accuracy

Funder, and others (Funder, 1995; Funder & Colvin, 1988; Funder & Sneed, 1993) presented the widely accepted golden standard for personality ground truth. This golden standard was based on the aggregation of the targets' self-ratings and their acquaintances' ratings of the targets themselves. Many research articles and papers produced for the psychology community have been based on this accuracy standard (Back & Nestler, 2016; Letzring, 2010; Naumann, Vazire, Rentfrow, & Gosling, 2009).

Moreover, Letzring, Wells, and Funder (2006) and Carney, Colvin, and Hall (2007) used the golden standard to determine how many seconds are required to make an accurate judgement of personality. Thin slices of time were used to make personality judgements. These papers identified important factors that affect the accuracy of a judgement. The authors stated that 60 seconds yielded enough information for a trait to be judged. In addition, the location, with regard to time, of the 60 seconds does not matter. Later, Krzyzaniak et al. (2019) studied the effect of several different thick slices of time, including 30 seconds, 1 minute, 3 minutes, and 5 minutes. Their results showed that 3 –4 minutes were sufficient to reveal enough information about one's personality to be judged. Moreover, 5 minutes did not add to the accuracy or quality of the information.

Accuracy and agreement of a personality judgement, are two different concepts. The former is the aim of personality research, while the latter is focused on the consensus of others regarding personality perception (Funder & West, 1993; Ready, Clark, Watson, & Westerhouse, 2000).

Accuracy requires the knowledge of the acquaintance to formulate a better understanding and accurate judgement of the judged target's personality (Colvin & Funder, 1991). Paunonen and Hong (2013) experimented with similarity ratings and indicated that acquaintances of the judged target can accurately rate the target's personality traits. Blackman and Funder (1998) found that less visible traits become more visible as the length of acquaintanceship increases. Dobewall, Aavik, Konstabel, Schwartz, and Realo (2014) emphasised the need for acquaintance-report to complement self-report to improve the accuracy of personality trait identification. Moreover, Vazire (2010) experimented with the self–other knowledge asymmetry model (SOKA). The author studied the agreement between strangers in comparison to the agreement between friends. The research showed that four strangers were less accurate than four friends in rating the target.

Hofstee (1994) challenged accuracy research by stating that others' reports of an individual's personality are more accurate than one's self-report. Several researchers tested this theory and concluded that having two acquaintances who know the target very well can result in a very accurate report, and two sources of information are better than one (Vazire & Mehl, 2008).

Kolar, Funder, and Colvin (1996) tested the theory of acquaintance and self-report with regard to accuracy, referencing the 'fish-and-water effect'. This refers to the fact that a fish loses awareness of its environment due to living there for a long time. The same is true for one's personality, as a person is unable to detect certain aspects of his personality because they have become too familiar and so he loses awareness of them. Another possibility was mentioned by Colvin, Block, and Funder (1995), who suggested that a person may report an inaccurate or distorted image of his personality. And in that case, it could be an overrated self-report. Indeed, people tend to think highly of themselves.

Epstein (1983) introduced two types of aggregation: appropriate and inappropriate aggregation. Appropriate aggregation reduces error variance when stimulus, situations, or judges are unrepresentative. Inappropriate aggregation can cause a loss of information, reliability, and validity.

Accuracy is an ongoing research field in the psychology community. The research covers many areas, including verbal and non-verbal cues, such as Facebook profiles, which include written text and profile pictures (Darbyshire, Kirk, Wall, & Kaye, 2016). Some research has been related to business, such as the accuracy of first impressions from employment interviews (Schmid Mast, Bangerter, Bulliard, & Aerni, 2011). One interesting paper examined the accuracy of teachers'ratings of their students (Dicke, Lüdtke, Trautwein, Nagy, & Nagy, 2012).

Research on accuracy in the psychology community is growing rapidly and covering many different aspects of life, work, and education (Beer, Rogers, & Letzring, 2019; Blackman & Funder, 1998; Hall, Gunnery, Letzring, Carney, & Colvin, 2017; Hall, Goh, Mast, & Hagedorn, 2016; Hall & Goh, 2017; Hirschmüller, Egloff, Schmukle, Nestler, & Back, 2015; Krzyzaniak & Letzring, 2019; Moritz & Roberts, 2018; Ter Laak, De Goede, & Brugman, 2001).

7.4 The Life Story Interview

Some psychology research has used the Life Story Interview as a stimulus to connect personality traits to certain questions (McAdams et al., 2004; McAdams, 2012).

The Life Story Interview serves as a stimulus for the three levels of personality (McAdams, 2001). Level one is dispositional traits. This includes the big five, which are stable and consistent behaviours throughout different situations. Level two is characteristic adaption. These are personal goals, adaptions, self-defence mechanisms, and coping strategies. Level three is life stories, which are the integrative part of personality that speaks to the identity of a person's self. The three levels form the complex human personality.

The life story has been used to study several factors related to human self-esteem, well-being, and more (Adler, Lodi-Smith, Philippe, & Houle, 2016). The life story has been used in part or full as a stimulus to reflect the big five personality traits (Bauer, McAdams, & Sakaeda, 2005b, 2005a; Raggatt, 2006; Thomsen, Olesen, Schnieber, & Tønnesvang, 2014; Coulter, Mallett, Singer, & Wrzus, 2018).

7.5 Realistic Accuracy Model (RAM)

Funder proposed the Realistic Accuracy Model (RAM), which was strongly influenced by Allport, Brunswick, and Gibson (Funder, 1995). To make an accurate judgement of personality, several elements must be present and in order. First, a stimulus is required to reveal the nature of the behaviour. Next, the behaviour is exhibited through different cues. Lastly, the perceiver or judge must detect those cues and interpret them correctly. This is the process of judgement proposed by Funder.

The RAM model states that there are four stages in accurately recognising a personality trait (Funder, 2012). The four stages are divided into two parts. Part one is the environment, which includes relevance and availability stages. Part two is the perceiver and includes detection and utilization. A failure in any stage or if used out of the designated order makes accuracy in personality identification impossible (Letzring et al., 2006; Letzring, Colman, Krzyzaniak, & Roberts, 2020).

The RAM is not dependent on personality trait cues alone. Figure 7.1 shows the RAM and its four stages. The stages are defined as follows:

1. Relevance stage: the target must present or emit a cue that is related to the trait to be judged.
2. Availability stage: this means the cue is available for the judge to rate. The cues can be verbal or non-verbal cues. However, thoughts or ideas in the target's brain, although they may be relevant, are not available to the judge.
3. Detection stage: is the responsibility of the judge to be able to detect the verbal or non-verbal cue only if it is relevant and available.
4. Utilization stage: this is the stage where the judge uses his ability to make a trait judgement from the relevant, available, and detected cues.

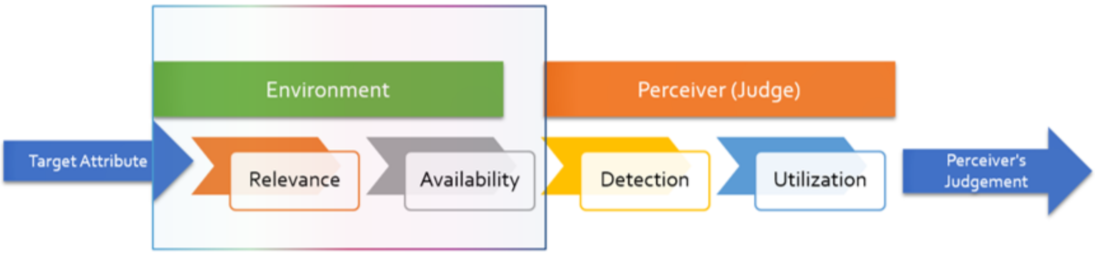


Figure 7.1: *Realistic Accuracy Model (RAM)(Funder, 1995)*

Chapter Summary

This chapter presented a foundation on which to build the remainder of this thesis. It showed that accuracy has not been captured fully in the computer science community. This may be the reason that research in computer science has slowed and is no longer being developed for commercial applications.

The aim of this chapter was to re-evaluate the research questions proposed in Chapter 1. The initial focus was on recognising honest signals and then developing a personality recognition model. However, with this revelation, new research questions were proposed (Chapter 1).

Personality recognition starts with accurate judgement of personality. In light of the research on personality psychology, building a personality recognition model starts with calculating an accurate personality trait ground truth. The revelations of accuracy of personality judgement, the RAM, and the need for a stimulus have shifted the focus to the gap and lack of guidelines or knowledge about personality psychology in the computer science community.

The major focus and goal have shifted to building a new personality corpus and answering the following research question: Can personality traits be recognised from non-verbal acoustic cues?

The following chapters are focused on collecting and testing the new personality traits corpus.

Chapter 8

The Personality Traits Corpus (PTC)

Research is creating new knowledge.

Neil Armstrong

This chapter describes in detail the data collection protocol for the new PTC. In addition, it describes the data within the corpus and how the ground truth was collected, calculated, and transformed. This study has received ethical approval from the University of Sheffield's ethics committee (ethics application no. 031314). All forms and sheets associated with the study can be found in Appendix B.

8.1 Design

This study required two types of participants: targets and their acquaintances. The target participants were recruited through Sheffield's University e-mail volunteer list and online websites, including social media outlets. Due to COVID-19 and lockdown restrictions, it was not possible to meet the targets face to face. An online registration form required interested target participants to provide the names and e-mail addresses of two acquaintances they have known for at least six months. The acquaintances were contacted by e-mail to complete the BFI-44 on the target. All participants (targets and acquaintances) received the information sheet and an online consent form that had to be accepted by the participants before proceeding with the study. Participants who did not consent were not allowed to participate in the study.

The targets provided at most four date and time options for the interview and survey. Then targets were e-mailed the details of the study and their confirmed interview invitation. On the selected date, the target received an e-mail with a link to the consent form and survey. The target could not proceed to the survey without providing consent to the study. After providing consent, the target could begin the BFI-44 survey about themselves. Once completed, the target logged into the online meeting room. The target could choose between Skype and Google Meet. When the target logged into the preferred meeting room, they were given the interview questions and instructions.

The targets were told the minimum time required for each question, not to mention people's names, only mention mild events regarding the low point question, and reminded that the interview would be audio recorded and would be stored in an online archive indefinitely. The targets were also reminded to turn off their camera. In addition, the targets were told that the recording would start after the main researcher read the first question and would end with the answer for question three. After each answer, when the target was silent, the researcher would ask the next question and so on. The result was one consecutive recording per target. At the end of the study, the main researcher asked the targets if they were okay. A list of mental health institutes, clinics, and health providers was prepared in case any participant required assistance. No target requested the list.

Due to COVID-19, one main advantage was that during the interview the target could not see the interviewer's face. This made the targets more comfortable because they could not see the main researcher's (interviewer) emotions or reactions to their answers. The target could answer without being judged. The main researcher spoke in a neutral tone and did not respond to the target's answers either verbally or non-verbally.

Editing Video

Skype and Google Meet do not specifically provide only audio recording. Therefore, all recordings were video recordings. However, the targets' and main researcher's cameras were turned off. All video recordings were converted to wave audio recordings.

The targets were instructed to give at least one-minute answers. Recordings varied in length between answers of the same target and between different targets.

Audio recordings were edited using Audacity[1] to include only the answers of the target to the three questions (positive childhood memory, mild low point, and turning point). Each edited recording started from the second the target answered and ended after 50–80 seconds to allow the target to complete a statement or a sentence. Total video lengths were between 3:06 and 4:03 minutes. If a participant discussed a traumatic event, their actual answer was not included; however, vague answers about the event were included. There were four videos that contained descriptions of very traumatic incidents, all of which were edited out.

8.2 Demographic Information

The survey included several questions to collect demographic data from both target and acquaintance participants. A short questionnaire was used to collect demographic information, including age, gender, ethnicity, nationality, and native language. In addition, acquaintances had to provide more information related to their relationship to the target, such as how long they have known them, how well they have known them, and their relationship to the target.

8.3 Personality Traits

The Big Five Inventory-44 (BFI-44) (John & Srivastava, 1999) measures five broad domains of personality: open-mindedness, conscientiousness, extraversion, agreeableness, and negative emotionality, with three facets per domain. The BFI-44 has 44 items rated on a Likert scale, ranging from 1 (disagree strongly) to 5 (agree strongly). The consent form, information sheet, interview questions, and BFI-44 questionnaires are presented in Appendix B.

8.4 Stimulus Material

Interview questions were selected from the Life Story Interview (Atkinson, 1998). The selected questions were closely connected with the big five personality traits and related to different events in a person's life. The questions were focused on a positive childhood memory, a mild low point, and a turning point. They are briefly described as follows:

[1] https://www.audacityteam.org/

1. **Positive childhood memory.** The fourth scene is an early memory – from childhood or your teen-aged years – that stands out as especially positive in some way. This would be a very positive, happy memory from your early years. Please describe this good memory in detail. What happened, where and when, who was involved, and what were you thinking and feeling? Also, what does this memory say about you or about your life?

2. **Low point.** The second scene is the opposite of the first. Thinking back over your entire life, please identify a scene that stands out as a low point, if not the low point in your life story. Even though this event is unpleasant, I would appreciate your providing as much detail as you can about it. What happened in the event, where and when, who was involved, and what were you thinking and feeling? Also, please say a word or two about why you think this particular moment was so bad and what the scene may say about you or your life. [Interviewer note: If the participant balks at doing this, tell him or her that the event does not really have to be the lowest point in the story but merely a very bad experience of some kind.]

3. **Turning point.** In looking back over your life, it may be possible to identify certain key moments that stand out as turning points – episodes that marked an important change in you or your life story. Please identify a particular episode in your life story that you now see as a turning point in your life. If you cannot identify a key turning point that stands out clearly, please describe some event in your life wherein you went through an important change of some kind. Again, for this event please describe what happened, where and when, who was involved, and what you were thinking and feeling. Also, please say a word or two about what you think this event says about you as a person or about your life.

8.5 Participants Data

As mentioned previously, there were two types of participants: targets and acquaintances. Targets were the main contributors, and acquaintances were secondary. Acquaintances complemented the targets'personality data.

8.5.1 Participants (Targets)

The participants for this study were recruited through Sheffield University's volunteer mailing list and a recruiting website (https://www.callforparticipants.com). The total number of participants was 86 (Figure 8.1). However, 10 participants did not show up for their scheduled interview meeting. Three participants withdrew from the study. Four participants did not reply back to confirm their appointments, and one participant was disqualified for ethical reasons. Of the 68 remaining participants who successfully completed the study, the mean age was 27.47 ($SD = 8.33$). The participants were 38% male and 62% female, as shown in Figure 8.2.

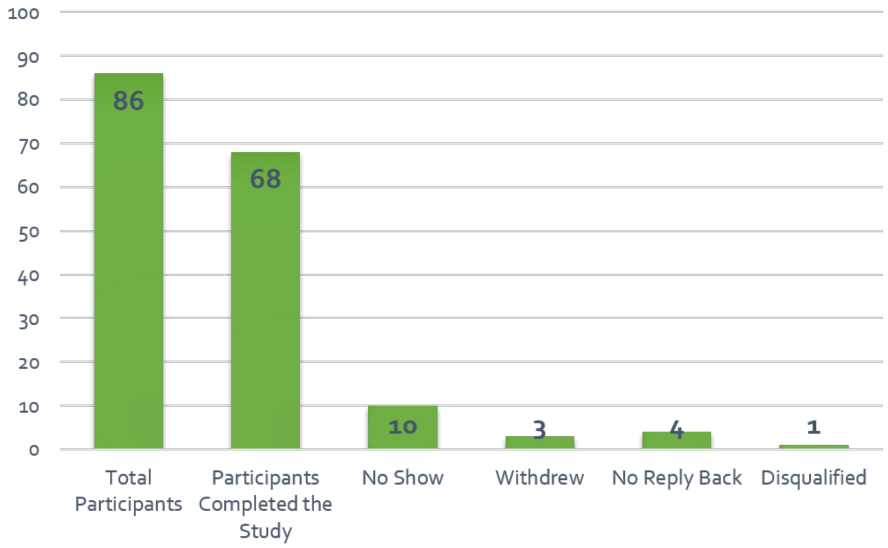


Figure 8.1: *Total number of target participants.*

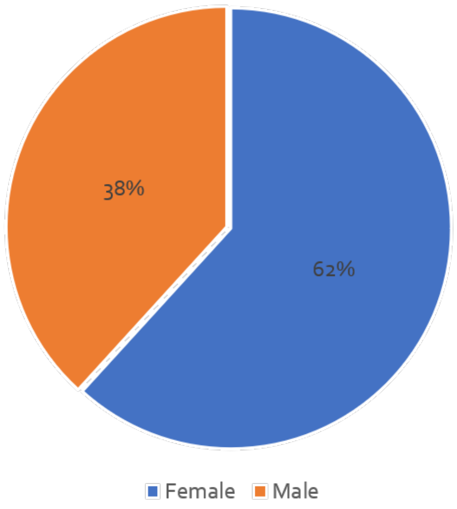


Figure 8.2: *Target participants' gender distribution*

The participants' ethnicities' varied: 58% white, 25% Asian, 7% Black/African American, 5% Latino/Hispanic/Spanish, 3% Arab, and 2% other ethnicities. In terms of nationality, the sample consisted of 50% British, 10% Romanian, 4% Indian, 3% of each of the following: Nigerian, Jordanian, Malaysian, and Singaporean. The remaining 24% was equally distributed between the following nationalities: Argentinian, Austrian, Cypriot, Ecuadorian, Estonian, Filipino, French, Greek, Irish, Italian, Mexican, Puerto Rican, Slovakian, South African, Turkish, and American.

Each target participant was required to provide the names and e-mails of two acquaintances who had known them for at least six months. The acquaintances were required to accept an online consent form before completing the online personality survey on the target participant.

8.5.2 Participants (Acquaintances)

The main researcher contacted 136 acquaintances. All acquaintances completed an online BFI-44 questionnaire about the target who nominated them. Their mean age was 31.24 ($SD = 12.07$). Of the acquaintances, 43% were male, and 57% were female. In terms of ethnicity, the majority, 58%, were white, 22% were Asians, 5% were Latino/Hispanic/Spanish, 4% were Arabs, 4% were Black/African American, and 7% were of other ethnicities. Regarding nationality, the majority, 55%, were British, 10% were Romanian, 3% were Malaysian, 2% were Estonian, 2% were Jordanian, 2% were Singaporean, and the remaining 26% were equally distributed between the following nationalities: Austrian, Chilean, Chinese, Greek, Irish, Italian, Turkish, American, Cameroonian, Colombian, Cypriot, Ecuadorian, French, Hong Kong, Nigerian, Polish, Puerto Rican, Salvadorian, Swiss, Slovakian, Thai, and Indian.

The acquaintances had known their targets for an average of 10.5 years ($SD = 9.15$). They had an average mean of 7.83 ($SD = 1.24$) for how well they knew them on a scale from 1 (note very well) to 9 (very well). The relationships between the target and acquaintances were collected and categorized as follows: 54% friends, 14% partners, 11% siblings, 11% mothers, 6% spouses, 2% colleagues, 1% fathers, and 1% relatives. The relationship types and distribution are shown in Figure 8.3

8.6 Ground Truth

From the BFI-44, the trait score was calculated for each target using BFI-44 questionnaire from the self-report, first acquaintance, and second acquaintance. The

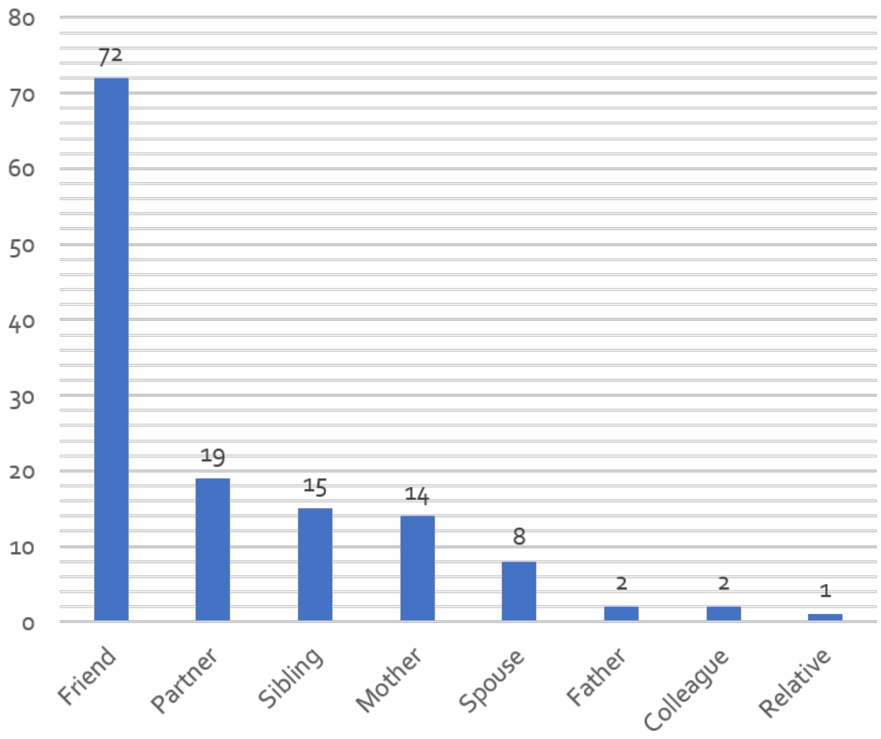
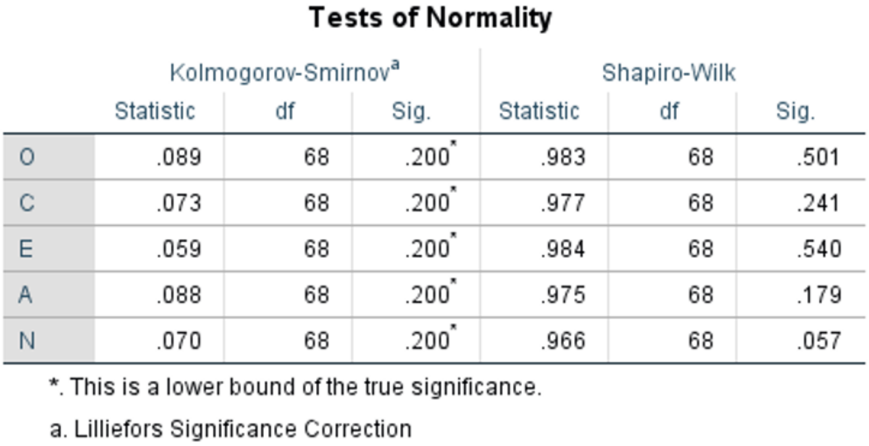


Figure 8.3: *Acquaintances'type of relationship type with the target.*

next step was to average the acquaintances' scores. Finally, the latter scores were averaged with the self-report score. The final number was the target's trait score. For each target, a score was calculated for each trait. Each target had five scores corresponding to each personality trait.

After the score for each trait per judge was calculated, it was transformed into z-score. A normality test was applied on scores for each trait. Figure 8.4 shows the results of both the Shapiro–Wilk and Kolmogorov–Smirnov normality tests. All ρ-values were above the threshold for significance. Hence, the null hypotheses for each trait was rejected, which states that the distribution was not a normal distribution. Neuroticism was slightly skewed. All normality distribution figures for the QQ plots and histograms are in Appendix C.

Personality scores were mostly classified as a high trait or a low trait. However, personality scores of average people can be close to the mean and therefore can be misclassified as high or low when in actuality it was an average score. Therefore, in this research and in accordance with personality psychology, personality scores were classified as high, medium, and low. In previous personality-related research, only a handful of studies used the three-level classification. It was apparent in those studies that there was a psychologist on the research team. Pianesi, Mana, Cappelletti,



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
O	.089	68	.200*	.983	68	.501
C	.073	68	.200*	.977	68	.241
E	.059	68	.200*	.984	68	.540
A	.088	68	.200*	.975	68	.179
N	.070	68	.200*	.966	68	.057

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

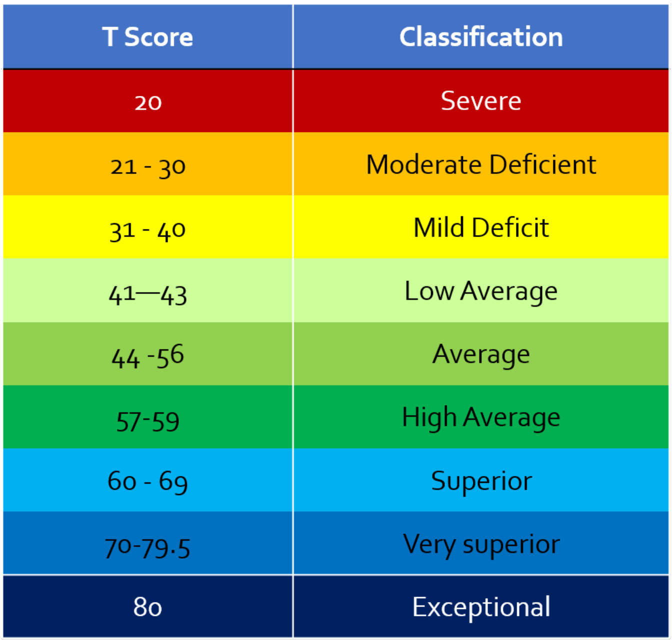
Figure 8.4: *Shapiro–Wilk and Kolmogorov–Smirnov tests of normality.*

Lepri, and Zancanaro (2008) used $1SD$ above and below the mean for the medium class. An and Levitan (2018) a used population norm threshold to classify their data into high, medium, and low. An et al. (2016) used the t-score to classify their data to three class labels.

Moreover, it is impossible to collect a sample that fully represents a population. According to Lew Goldberg (2006), "*One should be very wary of using canned* 'norms' *because it isn't obvious that one could ever find a population of which one's present sample is a representative subset. Most 'norms' are misleading, and therefore they should not be used*". Accordingly, the personality z-scores were transformed to t-scores, with $\mu = 50$ and $\sigma = 10$ (Equation 8.1). T-scores, are used in many personality and medical studies when it is difficult to capture the population mean. A conversion table was used to convert a t-score to a classification. Each t-score was associated with a class label. Figure 8.5 shows the conversion table for t-scores, which was derived from the original in Figure 8.6. It has nine score labels, from low to high: severe, moderate deficit, mild deficit, low average, average, high average, superior, very superior, and exceptional. In this research, every three labels were grouped to represent one label, as shown in Figure 8.7.

$$t = 50 + z(10) \tag{8.1}$$

Figure 8.8 shows the final count for each personality trait. It is clear that the sample majority was average for all personality traits. If the study only categorized a trait as high or low, a lot of the in-between values would have been misclassified,



T Score	Classification
20	Severe
21 - 30	Moderate Deficient
31 - 40	Mild Deficit
41—43	Low Average
44 -56	Average
57-59	High Average
60 - 69	Superior
70-79.5	Very superior
80	Exceptional

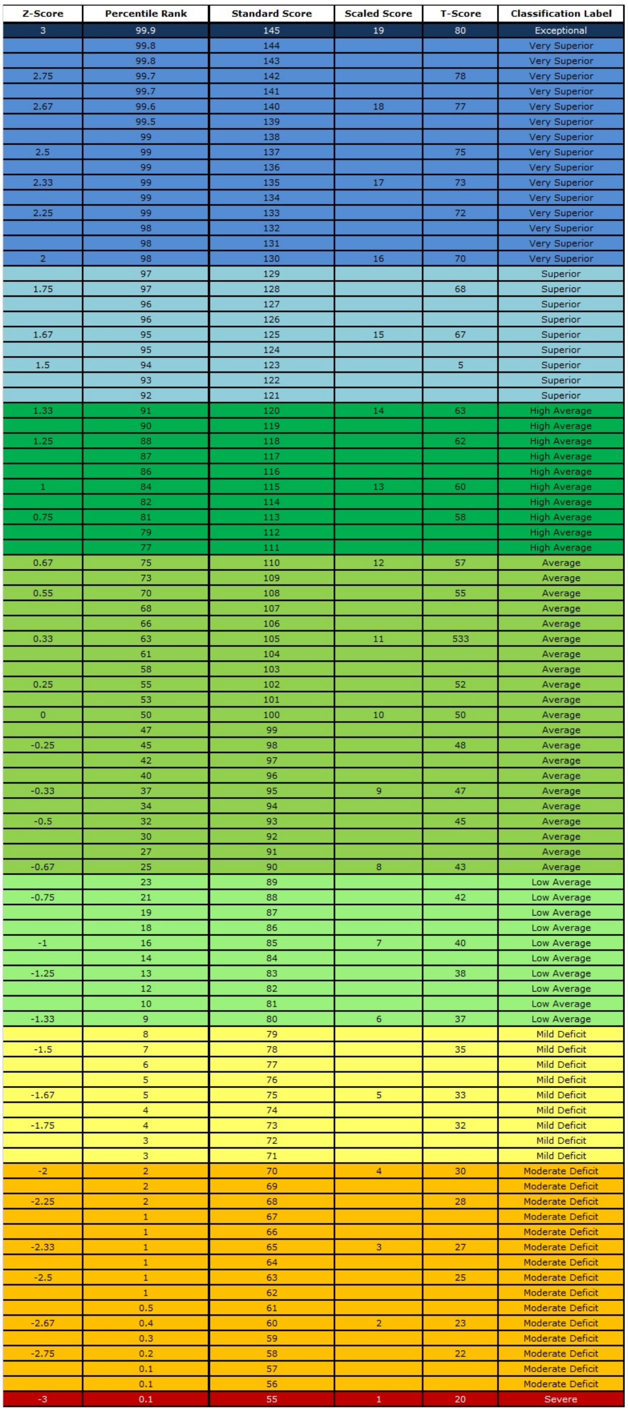
Figure 8.5: *T-score conversion table.*

and personality recognition would not have fulfilled its purpose.

The scores were calculated based on the acquaintance's relationship to the target, and the zero-acquaintance issue presented previously was resolved by including it in the final score. Therefore, this corpus has captured a more focused personality trait score than previous research in the area of personality recognition.

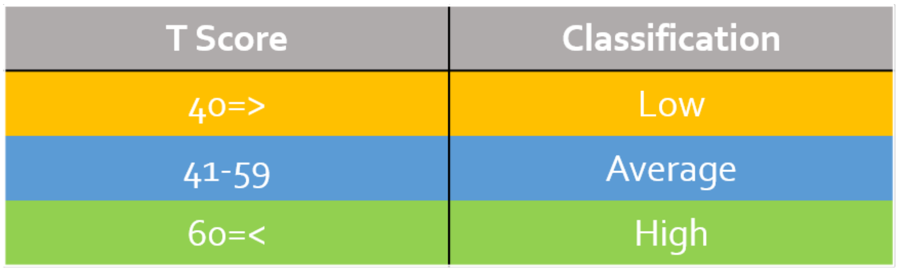
Cronbach's alpha for the self-report was $\alpha = 0.765$ and $\rho = 0.000$. Spearman's rho and Pearson's correlation were used to test the correlation between the self-report and each acquaintance and between the two acquaintances. The test was also performed between the self-report and acquaintances average. The correlation for openness between self-report and first acquaintance-report was $r = .537$ ($\rho = 0.000$). Meanwhile, the correlation between self-report and second acquaintance-report it was $r = 0.332$ ($\rho = 0.003$). The Pearson correlation was $r = 0.318$ ($\rho = 0.004$) between both acquaintances' reports.

Figure 8.9 shows the significant correlation between each trait of the self-report and the first and second acquaintance reports and between the two acquaintances reports. Figure 8.10 displays the Pearson correlations between the self-report score for each trait and the average trait score calculated from the two acquaintances' reports. It was clear that a significant correlation exists between the three raters' reports.



Z-Score	Percentile Rank	Standard Score	Scaled Score	T-Score	Classification Label
3	99.9	145	19	80	Exceptional
	99.8	144			Very Superior
	99.8	143			Very Superior
2.75	99.7	142		78	Very Superior
	99.7	141			Very Superior
2.67	99.6	140	18	77	Very Superior
	99.5	139			Very Superior
	99	138			Very Superior
2.5	99	137		75	Very Superior
	99	136			Very Superior
2.33	99	135	17	73	Very Superior
	99	134			Very Superior
2.25	99	133		72	Very Superior
	98	132			Very Superior
	98	131			Very Superior
2	98	130	16	70	Very Superior
	97	129			Superior
1.75	97	128		68	Superior
	96	127			Superior
	96	126			Superior
1.67	95	125	15	67	Superior
	95	124			Superior
1.5	94	123		5	Superior
	93	122			Superior
	92	121			Superior
1.33	91	120	14	63	High Average
	90	119			High Average
1.25	88	118		62	High Average
	87	117			High Average
	86	116			High Average
1	84	115	13	60	High Average
	82	114			High Average
0.75	81	113		58	High Average
	79	112			High Average
	77	111			High Average
0.67	75	110	12	57	Average
	73	109			Average
0.55	70	108		55	Average
	68	107			Average
	66	106			Average
0.33	63	105	11	533	Average
	61	104			Average
	58	103			Average
0.25	55	102		52	Average
	53	101			Average
0	50	100	10	50	Average
	47	99			Average
-0.25	45	98		48	Average
	42	97			Average
	40	96			Average
-0.33	37	95	9	47	Average
	34	94			Average
-0.5	32	93		45	Average
	30	92			Average
	27	91			Average
-0.67	25	90	8	43	Average
	23	89			Low Average
-0.75	21	88		42	Low Average
	19	87			Low Average
	18	86			Low Average
-1	16	85	7	40	Low Average
	14	84			Low Average
-1.25	13	83		38	Low Average
	12	82			Low Average
	10	81			Low Average
-1.33	9	80	6	37	Low Average
	8	79			Mild Deficit
-1.5	7	78		35	Mild Deficit
	6	77			Mild Deficit
	5	76			Mild Deficit
-1.67	5	75	5	33	Mild Deficit
	4	74			Mild Deficit
-1.75	4	73		32	Mild Deficit
	3	72			Mild Deficit
	3	71			Mild Deficit
-2	2	70	4	30	Moderate Deficit
	2	69			Moderate Deficit
-2.25	2	68		28	Moderate Deficit
	1	67			Moderate Deficit
	1	66			Moderate Deficit
-2.33	1	65	3	27	Moderate Deficit
	1	64			Moderate Deficit
-2.5	1	63		25	Moderate Deficit
	1	62			Moderate Deficit
	0.5	61			Moderate Deficit
-2.67	0.4	60	2	23	Moderate Deficit
	0.3	59			Moderate Deficit
-2.75	0.2	58		22	Moderate Deficit
	0.1	57			Moderate Deficit
	0.1	56			Moderate Deficit
-3	0.1	55	1	20	Severe

Figure 8.6: *T-score original conversion table.*



T Score	Classification
40=>	Low
41-59	Average
60=<	High

Figure 8.7: *T-score conversion table (Derived).*

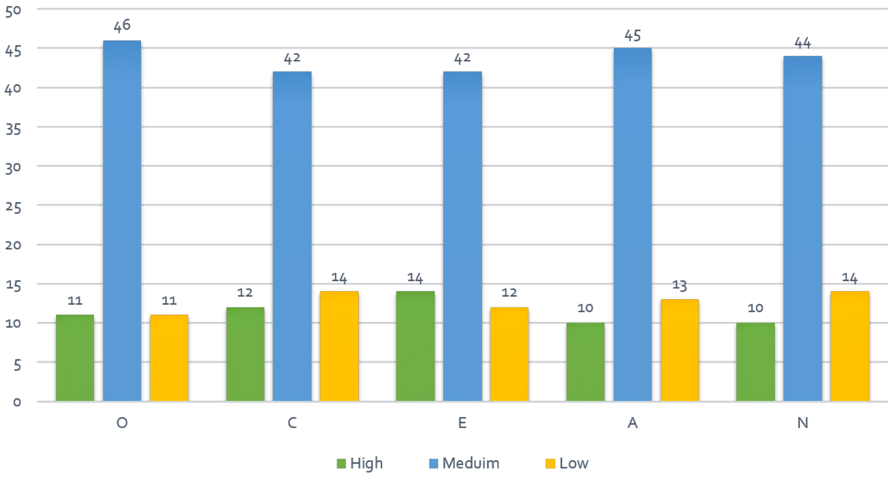
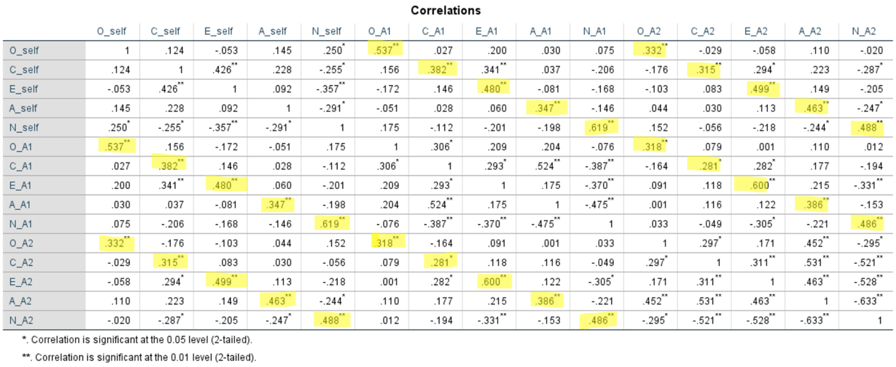


Figure 8.8: *Updated t-score conversion table.*



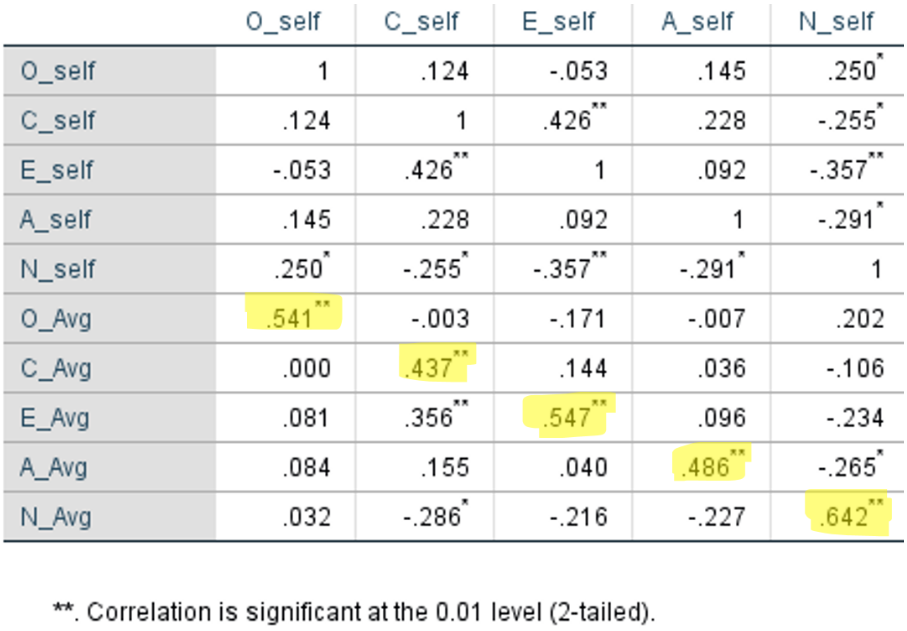
Correlations

	O_self	C_self	E_self	A_self	N_self	O_A1	C_A1	E_A1	A_A1	N_A1	O_A2	C_A2	E_A2	A_A2	N_A2
O_self	1	.124	-.053	.145	.250*	.537**	.027	.200	.030	.075	.332**	-.029	-.058	.110	-.020
C_self	.124	1	.426**	.228	-.255*	.156	.382**	.341**	.037	-.206	-.176	.315**	.294*	.223	-.287*
E_self	-.053	.426**	1	.092	-.357**	-.172	.146	.480**	-.081	-.168	-.103	.083	.499**	.149	-.205
A_self	.145	.228	.092	1	-.291*	-.051	.028	.060	.347**	-.146	.044	.030	.113	.463**	-.247*
N_self	.250*	-.255*	-.357**	-.291*	1	.175	-.112	-.201	-.198	.619**	.152	-.056	-.218	-.244*	.488**
O_A1	.537**	.156	-.172	-.051	.175	1	.306*	.209	.204	-.076	.318**	.079	.001	.110	.012
C_A1	.027	.382**	.146	.028	-.112	.306*	1	.293*	.524**	-.387**	-.164	.281*	.282*	.177	-.194
E_A1	.200	.341**	.480**	.060	-.201	.209	.293*	1	.175	-.370**	.091	.118	.600**	.215	-.331**
A_A1	.030	.037	-.081	.347**	-.198	.204	.524**	.175	1	-.475**	.001	.116	.122	.386**	-.153
N_A1	.075	-.206	-.168	-.146	.619**	-.076	-.387**	-.370**	-.475**	1	.033	-.049	-.305*	-.221	.486**
O_A2	.332**	-.176	-.103	.044	.152	.318**	-.164	.091	.001	.033	1	.297*	.171	.452**	-.295*
C_A2	-.029	.315**	.083	.030	-.056	.079	.281*	.118	.116	-.049	.297*	1	.311**	.531**	-.521**
E_A2	-.058	.294*	.499**	.113	-.218	.001	.282*	.600**	.122	-.305*	.171	.311**	1	.463**	-.528**
A_A2	.110	.223	.149	.463**	-.244*	.110	.177	.215	.386**	-.221	.452**	.531**	.463**	1	-.633**
N_A2	-.020	-.287*	-.205	-.247*	.488**	.012	-.194	-.331**	-.153	.486**	-.295*	-.521**	-.528**	-.633**	1

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 8.9: *Correlation scores between self-report, first acquaintance report, and second acquaintance report.*



	O_self	C_self	E_self	A_self	N_self
O_self	1	.124	-.053	.145	.250*
C_self	.124	1	.426**	.228	-.255*
E_self	-.053	.426**	1	.092	-.357**
A_self	.145	.228	.092	1	-.291*
N_self	.250*	-.255*	-.357**	-.291*	1
O_Avg	.541**	-.003	-.171	-.007	.202
C_Avg	.000	.437**	.144	.036	-.106
E_Avg	.081	.356**	.547**	.096	-.234
A_Avg	.084	.155	.040	.486**	-.265*
N_Avg	.032	-.286*	-.216	-.227	.642**

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 8.10: *Correlation scores between self-report and average acquaintances report.*

Chapter Summary

This chapter described the design of the personality study in detail, detailing the collection process step-by-step. Data were collected from each target participant and their two acquaintances using the BFI-44 questionnaire. Interview questions were selected from the Life Story Interview to elicit personality traits from the target participant. Exploratory data analysis was completed, and are visualised in figures in Appendix C. Personality trait scores were transformed to *t*-scores, as suggested in psychology research. Personality traits scores were classified into three categories: high, medium, and low. The PTC is ready to be used for experimentation in the next chapter. The first step toward answering the research questions has been completed. The next step is to apply machine learning algorithms on the PTC and determine if personality recognition is possible from non-verbal acoustic cues.
Figure 8.11 shows a flowchart of the process completed so far.

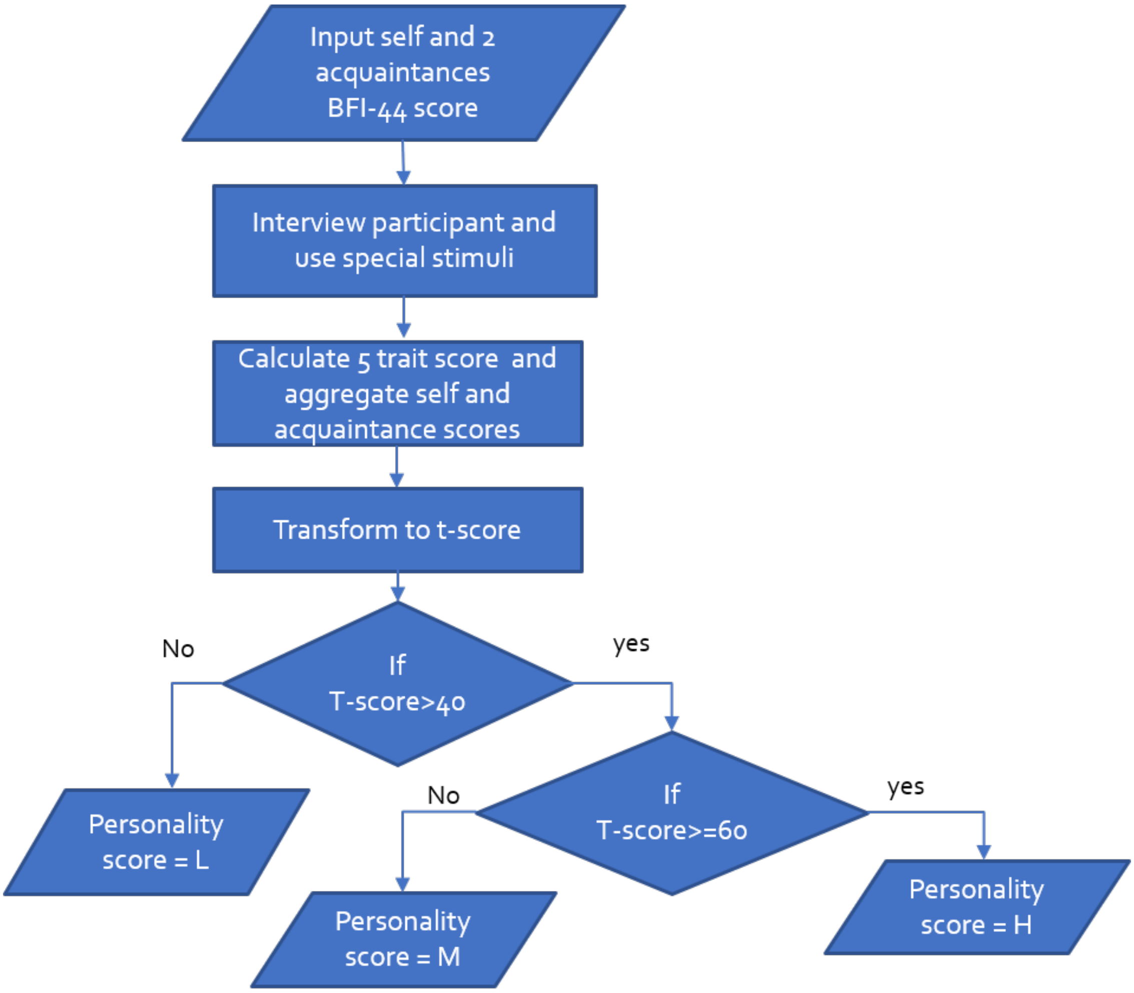

Input self and 2 acquaintances BFI-44 score
Interview participant and use special stimuli
Calculate 5 trait score and aggregate self and acquaintance scores
Transform to t-score
If T-score>40
No
yes
Personality score = L
If T-score>=60
No
yes
Personality score = M
Personality score = H

Figure 8.11: *Personality study protocol and ground-truth collection and calculation.*

Chapter 9

Experiments with the Personality Traits Corpus (PTC)

> *Curiosity - the rover and the concept - is what science is all about: the quest to reveal the unknown.*
>
> Ahmed Zewail

Data were classified in the previous chapter as high, medium, or low. OpenSMILE was used to extract acoustic features from each audio file. The configuration file used for feature extraction was compare2016 (Schuller et al., 2016). Over six thousand features were extracted. Demographic data were appended to the dataset to prepare the final personality recognition corpus. In addition, four attributes were added to the dataset: interview duration, question one's answer duration in seconds, question two's answer duration in seconds, and question three's answer duration in seconds.

The amount of collected data is relatively small when compared with the speaker personality corpus and other corpora. The huge number of features compared to the number of instances indicate that this corpus is a high-dimensionality dataset. Before applying any feature reduction techniques, the corpus was used to build eighteen classification models. All of the classification models were built with default settings, and no parameter tuning was applied. No classifier settings were altered at this stage.

Personality recognition is still a new field, and therefore multiple classifiers must be tested on this type of problem to select the best-performing classifier or classifiers.

The aim was to find the best classifier(s) for the PTC. This was a multi-class problem, and thus the recall and AUC baselines were 33% and 50%, respectively.

Therefore, eighteen classifiers were trained and tested on the PTC. The different classifiers included: DL, SGD, LR, RF, Adaboost, SVM and SVC-Linear, κNN, bagging, PA, RC, ANN, perceptron, GB, LDA, QDA, naïve Bayes, and DTC.

9.1 Experiment 1: Default Setting

In the first experiment with the new corpus, the full dataset was used to build all classifiers with cross-validation $\kappa = 10$, and no parameter tuning was applied. The first experiment was completed without any significant results from any classifier and across all traits. Therefore, MCC was used as a measure of classifier statistical significance. The closer the value is to $+1$, the higher the agreement between predictions and actual values. Furthermore, accuracy has been shown to be an inappropriate and misleading performance measure.

Table 9.1 shows the classifiers' performance and their ability to predict the openness trait. The best-performing classifier was naïve Bayes, with a 59% UAR and supported with MCC at 0.40, which indicates that the classifier was not predicting randomly. In addition, the AUC value supports the classifier's ability to distinguish between true positives and true negatives (69%). The highest accuracy for the openness trait was achieved with the naïve Bayes algorithm; however, at 71% it was evident that it was not a clear reflection of the classifier's accuracy. This was caused by classifying most instances as majority class. Therefore, recall is a more appropriate measure.

From Table 9.2 it was evident that, for the conscientiousness trait, decision tree, and SVM algorithms produced the highest accuracy at 61%. However, this is a misleading evaluation measure, and UAR was preferred. The conscientiousness trait was the worst-performing trait, and its recall for two classification algorithms, ANN and ridge, was 33%. Furthermore, AUC and MCC measures indicated the randomness of the classification algorithms' performance. The AUC of 50% and MCC below 0.10 both show clear evidence of prediction randomness and a lack of significance.

Table 9.3 shows that κNN was the best-performing classification algorithm for the extraversion trait, with a 53% recall, 66% AUC, and an MCC significance of 0.36. Coincidently, accuracy was also highest for κNN and logistic regression. However, it

Table 9.1: *Openness trait UAR, AUC, accuracy, and MCC.*

Classifier	Accuracy	UAR	AUC	MCC
Decision Tree	47.62%	29.76%	48.65%	0.00
Perceptron	33.33%	46.03%	59.78%	0.18
ANN	66.67%	50.79%	63.04%	0.28
Deep Learning	57.14%	43.25%	58.76%	0.18
SVM	66.67%	33.33%	50.00%	None
Naïve Bayes	**71.43%**	**59.13%**	**69.58%**	**0.40**
Logistic Regression	66.67%	48.02%	63.05%	0.30
κNN	57.14%	37.30%	52.98%	0.06
Bagging	61.90%	30.95%	47.82%	-0.11
Random Forest	66.67%	33.33%	50.00%	None
Adaboost	66.67%	42.06%	57.27%	0.21
Linear SVC	28.57%	43.65%	57.60%	0.14
Passive Aggressive	33.33%	43.25%	58.33%	0.17
Ridge	52.38%	34.92%	50.87%	0.00
Gradient Boosting	57.14%	34.52%	50.08%	0.00
LDA	52.38%	26.19%	43.59%	-0.19
SGD	42.86%	48.02%	62.62%	0.25
QDA	28.57%	20.24%	44.55%	-0.03

is not an acceptable measure for this case.

The agreeableness trait fared similarly to the conscientiousness trait, as shown in Table 9.4. Gradient boosting achieved the highest recall at 45% and was supported by weak measure of MCC at 0.15. The AUC was 57%, further indicating the weak predictive capability of the classification algorithm. For the agreeableness trait, accuracy was 66% with SVM, but its MCC produced a result indivisible by zero, signifying the lack of significance in any of the classifier's predictions'.

In Table 9.5, neuroticism is another trait with very low performance scores. κNN's recall was 47% and its AUC score of 60% indicated that the classification algorithm was weakly skilled at prediction. MCC suggested that the classifier's performance was significant and that it was not randomly predicting instances. Similar to the openness trait, accuracy was high at 71%, but this may be misleading because, due to the majority class, most instances were predicted as majority instances.

In the previous section, it was shown that accuracy as a measure can be inappropriate and misleading. This is caused during the training phase of the experiment. During training, regardless of cross-validation, the classification algorithms were exposed more to the majority class, which eventually appeared in each cross-validation split. This caused the algorithms to learn more about the majority

Table 9.2: *Conscientiousness trait UAR, AUC, accuracy, and MCC.*

Classifier	Accuracy	UAR	AUC	MCC
Decision Tree	**61.90%**	33.33%	53.30%	0.16
Perceptron	14.29%	19.23%	40.86%	-0.15
ANN	52.38%	**33.97%**	**50.56%**	0.01
Deep Learning	33.33%	35.26%	48.38%	-0.06
SVM	**61.90%**	33.33%	50.00%	None
Naïve Bayes	57.14%	30.77%	48.84%	-0.02
Logistic Regression	52.38%	33.97%	49.46%	-0.03
κNN	38.10%	26.28%	42.67%	-0.18
Bagging	42.86%	28.85%	46.04%	-0.09
Random Forest	42.86%	23.08%	40.95%	-0.26
Adaboost	57.14%	30.77%	48.84%	-0.02
Linear SVC	28.57%	32.69%	50.53%	0.02
Passive Aggressive	23.81%	30.13%	48.27%	-0.02
Ridge	52.38%	**33.97%**	**50.56%**	0.01
Gradient Boosting	52.38%	28.21%	45.47%	-0.17
LDA	52.38%	28.21%	45.47%	-0.17
SGD	14.29%	13.46%	36.87%	-0.22
QDA	23.81%	24.36%	45.39%	-0.06

class, and no or very little training from minority classes was achieved. When the testing phase was conducted, the classification algorithms incorrectly predicted most instances as a majority class label. This was clearly evident in the discrepancies between the accuracy and recall scores and further supported by the AUC and MCC.

Across all traits, the highest recall score for each trait was supported by the highest AUC and the highest positive MCC score. This suggests that, despite the low performance, the selected classification algorithms were functioning properly and not with complete randomness. Further supporting figures and detailed tables are provided in Appendix D.

9.2 Experiment 2: Hyperparameter Tuning

In this experiment and the experiments that followed, three-way holdout method was applied. Figure 9.1 shows the process of how the dataset was split and the models were built and tested. The test set was not part of the training to avoid over-fitting and yield a better generalizable model.

The dataset was split into training (70%) and testing (30%) sets. The training

Table 9.3: *Extraversion trait UAR, AUC, accuracy, and MCC.*

Classifier	Accuracy	UAR	AUC	MCC
Decision Tree	61.90%	**56.41%**	**65.95%**	0.30
Perceptron	38.10%	43.59%	57.94%	0.14
ANN	61.90%	50.64%	61.96%	0.25
Deep Learning	38.10%	37.82%	51.75%	0.01
SVM	61.90%	33.33%	50.00%	None
Naïve Bayes	52.38%	33.97%	49.46%	-0.03
Logistic Regression	**66.67%**	47.44%	60.23%	0.29
κNN	**66.67%**	53.21%	66.43%	0.36
Bagging	61.90%	44.87%	57.97%	0.19
Random Forest	61.90%	33.33%	50.00%	None
Adaboost	42.86%	28.85%	48.24%	-0.01
Linear SVC	28.57%	32.69%	49.43%	-0.01
Passive Aggressive	28.57%	32.69%	49.43%	-0.01
Ridge	61.90%	44.87%	57.97%	0.19
Gradient Boosting	42.86%	23.08%	40.95%	-0.25
LDA	57.14%	36.54%	51.72%	0.03
SGD	28.57%	32.69%	50.53%	0.02
QDA	52.38%	33.97%	50.56%	0.01

set was further split, with 70% for training and 30% for validation. However, the aim of a dedicated test set is to maintain its concealed status and not be part of or leak into the training set. Therefore, a fixed number for test set was selected, which was a dedicated test set throughout the remainder of the experiments. Table 9.6 shows the distribution of class labels per trait per test set.

This experiment adopted the dedicated test set, which remained unknown during the building and evaluating phases. The test set is the true measure of the classification algorithm's performance and is more generalizable.

It is clear in Table 9.7 that the openness trait was hard to predict. Most classification algorithms were unsuccessful and random. Perceptron was the only classifier with a 43% UAR and AUC of 57%, which indicated its prediction was not completely random. MCC was over zero, which signified the perceptron's slight ability predicting the correct class label.

The evaluation results of the conscientiousness trait after hyperparameter tuning are presented in Table 9.8. Decision tree was the best-performing classification algorithm, with a UAR of 39%. The AUC and MCC were slightly higher, indicating non-random prediction.

Table 9.4: *Agreeableness trait UAR, AUC, accuracy, and MCC.*

Classifier	Accuracy	UAR	AUC	MCC
Decision Tree	38.10%	27.78%	42.95%	-0.17
Perceptron	14.29%	15.87%	37.86%	-0.21
ANN	23.81%	11.90%	31.86%	-0.40
Deep Learning	23.81%	20.63%	40.70%	-0.16
SVM	**66.67%**	33.33%	50.00%	None
Naïve Bayes	57.14%	28.57%	47.05%	-0.07
Logistic Regression	47.62%	23.81%	41.37%	-0.24
κNN	38.10%	19.05%	38.42%	-0.25
Bagging	28.57%	14.29%	32.68%	-0.39
Random Forest	57.14%	28.57%	47.05%	-0.07
Adaboost	52.38%	26.19%	43.59%	-0.19
Linear SVC	19.05%	24.21%	44.41%	-0.09
Passive Aggressive	19.05%	24.21%	44.41%	-0.09
Ridge	57.14%	28.57%	45.71%	-0.16
Gradient Boosting	61.90%	**45.63%**	**57.97%**	0.15
LDA	52.38%	26.19%	46.34%	-0.06
SGD	14.29%	15.87%	37.86%	-0.21
QDA	38.10%	36.51%	50.17%	-0.03

Extraversion performed better than previous traits with hyperparameter tuning. Passive aggressive had the highest evaluation measure across all algorithms. Its AUC and MCC scores were 63% and 0.27, respectively. This implied that the classifier was performing above random predictions and had the ability to predict the actual class, although not often, because MCC was closer to zero than to +1. This is shown in Table 9.9.

Table 9.10 presents the evaluation measure of the agreeableness trait. Its performance was similar to that of the previous traits except extraversion. Deep learning was the best-performing classification algorithm, with a UAR of 41%; furthermore, its AUC and MCC scores were slightly above the level of randomness.

Lastly, neuroticism performance is shown in Table 9.11. ANN was its best-performing classifier, with a UAR of 39%. Like agreeableness, its AUC and MCC scores were slightly above the level of randomness.

Experiment 2 was different because there were three sets derived from the full dataset: the training set, validation set, and a test set. Therefore, the training and validation sets were reduced and contained a few minority class samples. The majority class was not affected because it was represented in all sets, and therefore the algorithm models were trained to predict it. Furthermore, if the focus shifted

Table 9.5: *Neuroticism trait UAR, AUC, accuracy, and MCC.*

Classifier	**Accuracy**	**UAR**	**AUC**	**MCC**
Decision Tree	47.62%	29.76%	48.49%	-0.01
Perceptron	38.10%	39.68%	53.00%	0.02
ANN	52.38%	26.19%	44.88%	-0.13
Deep Learning	42.86%	27.38%	45.08%	-0.09
SVM	66.67%	33.33%	50.00%	None
Naïve Bayes	42.86%	21.43%	40.65%	-0.20
Logistic Regression	57.14%	28.57%	47.05%	-0.07
κNN	**71.43%**	**47.62%**	**60.92%**	0.33
Bagging	61.90%	36.90%	52.20%	0.05
Random Forest	66.67%	33.33%	50.00%	None
Adaboost	28.57%	20.24%	37.06%	-0.29
Linear SVC	33.33%	25.40%	43.58%	-0.12
Passive Aggressive	33.33%	31.35%	47.96%	-0.04
Ridge	52.38%	26.19%	43.54%	-0.20
Gradient Boosting	42.86%	21.43%	40.65%	-0.20
LDA	0.62%	30.95%	47.82%	-0.11
SGD	33.33%	31.35%	46.51%	-0.09
QDA	14.29%	21.83%	41.17%	-0.16

Table 9.6: *Test set class distribution and total for each personality trait.*

Trait\Labels	**High**	**Medium**	**Low**	**Test Set Total**
Openness	10	14	10	**34**
Conscientiousness	11	13	12	**36**
Extraversion	12	13	11	**36**
Agreeableness	9	14	12	**35**
Neuroticism	9	13	13	**35**

from macro recall (UAR) to recall of the majority class, then the score would be over 70%. This could be very misleading if only majority class recall was reported. The focus was to train and build classifiers with the ability to predict all personality traits' classes labels and not a single class label.

Regarding QDA, it fails to perform if the training samples from a minority class are very low. Therefore, in this experiment QDA produced no scores for any personality trait.

It is noticeable that the best-performing classifier for each personality trait had the lowest ρ-value and a positive MCC score.

Interestingly, it is apparent from this experiment that not a single classifier could

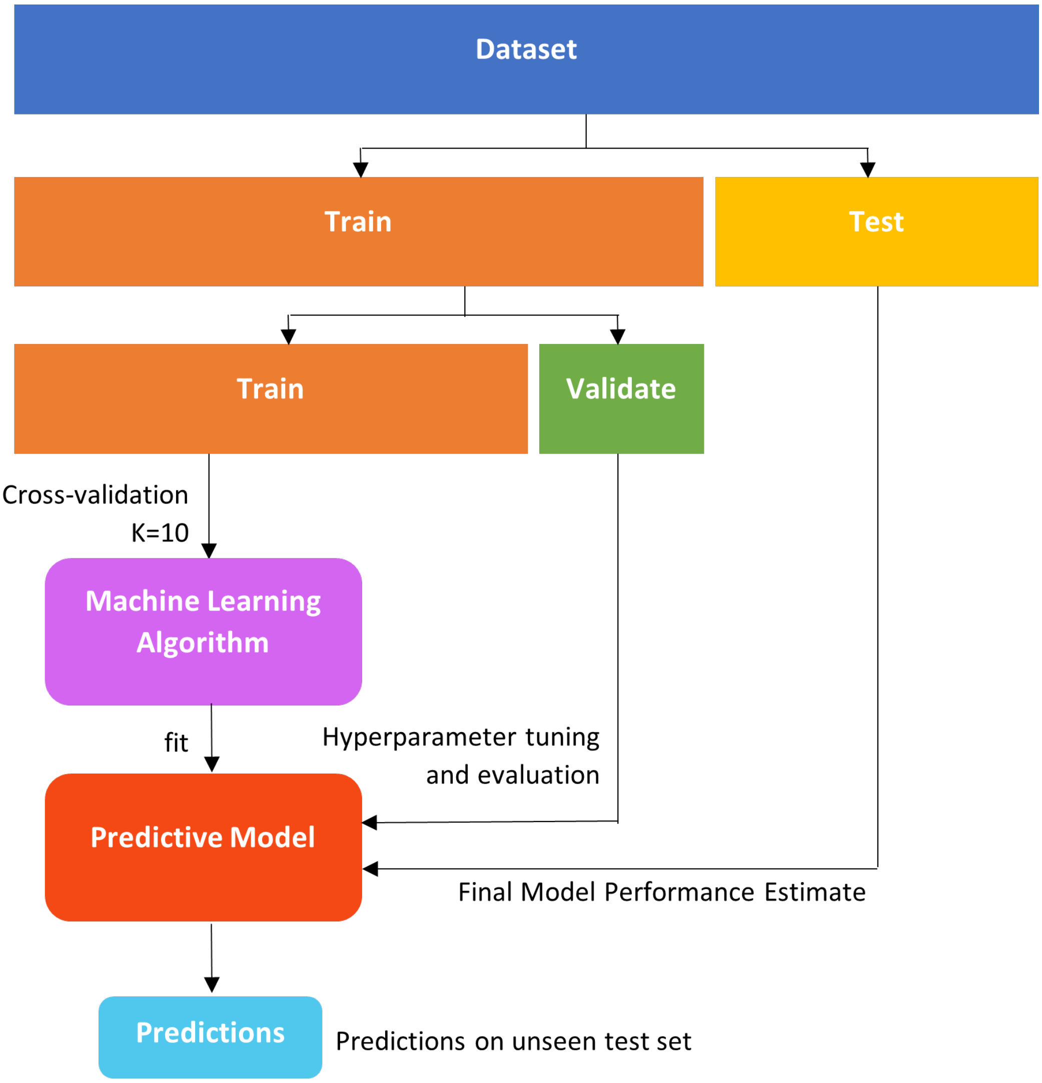

Dataset
Train
Test
Train
Validate
Cross-validation
K=10
Machine Learning Algorithm
fit
Hyperparameter tuning and evaluation
Predictive Model
Final Model Performance Estimate
Predictions
Predictions on unseen test set

Figure 9.1: *The proposed framework model for personality recognition with three-way hold-out method.*

predict all personality traits. For each personality trait, one classifier outperformed the rest. This will be further investigated in the next section.

9.3 Experiment 3: Corpus Augmentation

Experiment three focused on the data imbalance issue that affected the model training step during model building. The model was well trained on the majority class and not enough on minority classes. Therefore, the next step involved balancing the dataset so the model could be trained and built with an equal training opportunity for all classes. The following sections explore data augmentation and the results

Table 9.7: *Openness trait evaluation measure with hyperparameter tuning.*

Classifier	Accuracy	UA Precision	UA Recall	UA F1	AUC	MCC	ρ-Value
Decision Tree	38.24%	13.13%	30.95%	18.44%	48.12%	-0.13	0.696
Perceptron	44.12%	**43.70%**	**43.33%**	**43.46%**	57.36%	**0.14**	0.427
ANN	41.18%	14.14%	33.33%	19.86%	50.14%	0.017	0.565
Deep Learning	35.29%	35.39%	33.33%	33.03%	50.00%	0.00	0.807
SVM	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Naïve Bayes	41.18%	14.14%	33.33%	19.86%	50.14%	0.01	0.565
Logistic Regression	44.12%	47.47%	36.67%	25.92%	52.50%	0.16	0.427
κNN	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Bagging	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Random Forest	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Adaboost	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Linear SVC	38.24%	37.08%	36.67%	36.67%	52.50%	0.05	0.696
Passive Aggressive	29.41%	29.04%	28.57%	28.69%	46.65%	-0.06	0.944
Ridge	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
Gradient Boosting	41.18%	13.73%	33.33%	19.44%	50.00%	None	0.565
LDA	23.53%	8.60%	26.67%	13.01%	45.28%	-0.18	0.990
QDA							
SGD	35.29%	34.92%	34.29%	34.43%	50.62%	0.01	0.807

Table 9.8: *Conscientiousness trait evaluation measure with hyperparameter tuning.*

Classifier	Accuracy	UA Precision	UA Recall	UA F1	AUC	MCC	ρ-Value
Decision Tree	41.67%	28.40%	**39.32%**	**31.03%**	**54.59%**	**0.12**	0.297
Perceptron	25.00%	18.75%	24.94%	21.37%	43.72%	-0.13	0.944
ANN	36.11%	23.23%	33.80%	22.15%	50.35%	0.01	0.562
Deep Learning	38.89%	29.80%	36.11%	23.60%	52.20%	0.09	0.425
SVM	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Naïve Bayes	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Logistic Regression	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
κNN	30.56%	16.67%	28.67%	19.43%	46.51%	-0.10	0.805
Bagging	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Random Forest	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Adaboost	38.89%	46.08%	36.11%	23.57%	52.17%	0.11	0.425
Linear SVC	27.78%	20.22%	27.97%	23.40%	45.93%	-0.08	0.889
Passive Aggressive	30.56%	30.56%	30.03%	29.01%	47.54%	-0.05	0.805
Ridge	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Gradient Boosting	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
LDA	36.11%	12.38%	33.33%	18.06%	50.03%	0.00	0.562
QDA							
SGD	27.78%	20.22%	27.97%	23.40%	45.93%	-0.08	0.889

achieved by augmenting the dataset.

9.3.1 Data Augmentation

Data augmentation is the process of increasing or augmenting observed data to make the fit better and the data analysis easier. It is simply observed data y augmented by quantity z, which is known as the latent data (Tanner & Wong, 1987). The augmentation scheme aims to increase the amount of data available to train machine learning algorithms. Data augmentation was first used to increase training data by Yaeger, Lyon, and Webb (1997). The word augmentation was not used to describe it, rather, they used the term stroke warping. Data augmentation differs from other sampling techniques. The sampling techniques introduced earlier were not a good

Table 9.9: *Extraversion trait evaluation measure with hyperparameter tuning.*

Classifier	Accuracy	UA Precision	UA Recall	UA F1	AUC	MCC	ρ-Value
Decision Tree	30.56%	11.46%	28.21%	16.30%	46.11%	-0.14	0.805
Perceptron	41.67%	46.57%	41.61%	42.06%	56.41%	0.13	0.297
ANN	27.78%	10.10%	25.64%	14.49%	44.15%	-0.24	0.889
Deep Learning	36.11%	32.26%	34.48%	29.10%	50.90%	0.02	0.562
SVM	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Naïve Bayes	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Logistic Regression	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
κNN	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Bagging	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Random Forest	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Adaboost	30.56%	10.78%	28.21%	15.60%	46.10%	-0.19	0.805
Linear SVC	41.67%	46.67%	41.61%	42.15%	56.44%	0.13	0.297
Passive Aggressive	50.00%	**53.47%**	**50.91%**	**49.27%**	**63.15%**	0.27	0.061
Ridge	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
Gradient Boosting	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
LDA	36.11%	12.04%	33.33%	17.69%	50.00%	None	0.562
QDA							
SGD	44.44%	46.99%	44.64%	44.58%	58.56%	0.17	0.191

Table 9.10: *Agreeableness trait evaluation measure with hyperparameter tuning.*

Classifier	Accuracy	UA Precision	UA Recall	UA F1	AUC	MCC	ρ-Value
Decision Tree	40.00%	22.31%	33.73%	23.43%	50.41%	0.01	0.563
Perceptron	31.43%	30.64%	33.47%	31.48%	49.44%	-0.02	0.887
ANN	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Deep Learning	48.57%	**48.39%**	**41.67%**	**34.07%**	**56.70%**	0.25	0.193
SVM	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Naïve Bayes	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Logistic Regression	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
κNN	37.14%	13.13%	30.95%	18.44%	48.32%	-0.07	0.694
Bagging	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Random Forest	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Adaboost	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Linear SVC	34.29%	34.03%	35.85%	34.49%	51.36%	0.01	0.804
Passive Aggressive	31.43%	30.31%	30.95%	30.04%	48.37%	-0.03	0.887
Ridge	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
Gradient Boosting	40.00%	13.33%	33.33%	19.05%	50.00%	None	0.563
LDA	40.00%	30.21%	34.66%	24.90%	50.88%	0.03	0.563
QDA							
SGD	40.00%	40.00%	41.01%	39.96%	55.37%	0.09	0.563

option in this scenario for the following reasons (Fernández, García, Herrera, & Chawla, 2018):

1. High-dimensional datasets: The large numbers of features lead to difficulties in building the model. A major issue is that machine learning algorithms consider all features when building the ideal model. Another major issue is the large number of features, which can overlap when seeking classification, resulting in overfitting.

2. Separability: This is considered very damaging to the data model training process. A model performs worst when it fails to clearly separate the classes. This is affected by the large number of features.

By creating synthetic data, more problems and overlapping are presented to the

Table 9.11: *Neuroticism trait evaluation measure with hyperparameter tuning.*

Classifier	Accuracy	UA Precision	UA Recall	UA F1	AUC	MCC	ρ-Value
Decision Tree	34.29%	11.76%	30.77%	17.02%	47.96%	-0.14	0.695
Perceptron	31.43%	31.00%	33.90%	31.82%	49.70%	-0.02	0.807
ANN	42.86%	**63.54%**	**39.60%**	30.08%	**54.77%**	0.19	0.296
Deep Learning	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
SVM	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Naïve Bayes	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Logistic Regression	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
κNN	34.29%	22.57%	30.77%	20.46%	47.96%	-0.08	0.695
Bagging	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Random Forest	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Adaboost	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Linear SVC	31.43%	31.00%	33.90%	31.52%	49.82%	-0.01	0.807
Passive Aggressive	34.29%	34.17%	37.61%	34.19%	52.43%	0.03	0.695
Ridge	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
Gradient Boosting	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
LDA	37.14%	12.38%	33.33%	18.06%	50.00%	None	0.563
QDA							
SGD	25.71%	23.94%	28.77%	25.52%	45.74%	-0.10	0.945

system. This has led to the avoidance of SMOTE techniques for data augmentation.

Further research in the area of data size and model performance was presented by Banko and Brill (2001) and Simard et al. (2003). Data augmentation can be performed in data space or feature space (Wong, Gatt, Stamatescu, & McDonnell, 2016). The former involves augmenting the data itself, while the latter augments the data's features. Sometime, not all features are presented or selected.

Image data augmentation techniques involve basic image manipulation (Shorten & Khoshgoftaar, 2019), such as flipping the image, changing its colour, cropping, rotation, shifting, and noise injection. Figure 9.2 shows an original image and several augmented images created by applying basic manipulation techniques. Further research on data augmentation included speech augmentation (Ragni, Knill, Rath, & Gales, 2014). Recent research interest has shifted to automatic speech recognition (ASR) data augmentation.

Audio manipulation techniques can also be used, such as adding noise and changing the pitch and speed. Figures 9.3 and 9.4 show example of pitch and speed changes for audio augmentation.

The new dataset was clearly unbalanced. However, there is no agreement in the research community regarding what is considered an imbalanced dataset. Some published research (Fernández, López, Galar, Del Jesus, & Herrera, 2013; Lango, 2019) suggested that for a dataset to be imbalanced, the imbalance ratio (IR) must be greater than 1.5. The imbalance ratio is calculated as one vs one (OVO). Table 9.12 shows the distribution of the class labels and the IR scores of the majority and each minority class.

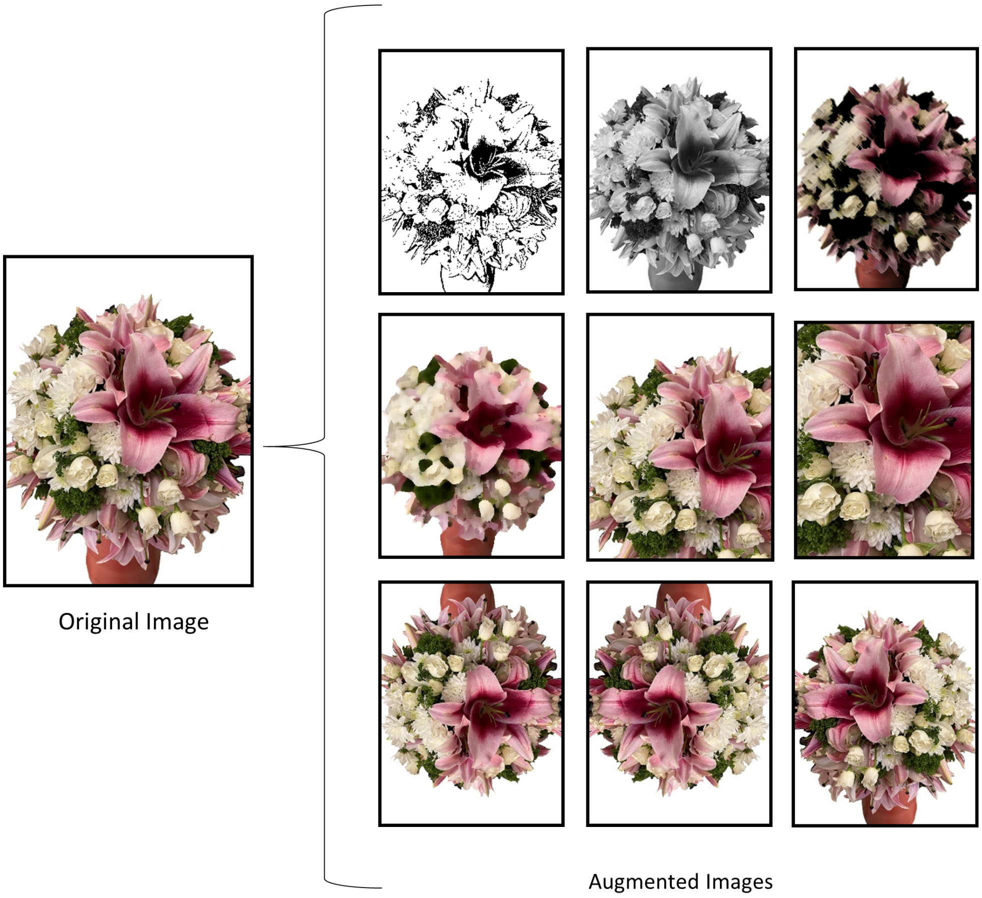


Figure 9.2: *Image augmentation samples.*

9.3.1.1 SpecAugment

Recent research on audio augmentation has introduced SpecAugment (Park et al., 2019). The aim was to augment speech data for several speech datasets, including Listen, Attend and Spell. The suggested technique was based on augmenting the data using mel spectrograms. The idea was to augment the input data. The authors wrote four different policies based on two speech datasets: LibriSpeech and Switchboard.

LibriSpeech (Panayotov, Chen, Povey, & Khudanpur, 2015) was first introduced for ASR tasks. The audio tracks were read in English from audio books from the LibriVox project. It had over 1000 hours of speech and was recorded at 16 KHz. Each clip was a monologue by a single speaker. All test clips were between 8 and 10 minutes long, while the training clips were 25 to 30 minutes long. The total number of speakers was 2484 (52% male, 48% female).

Switchboard (Godfrey, Holliman, & McDaniel, 1992) was presented in 1998. The

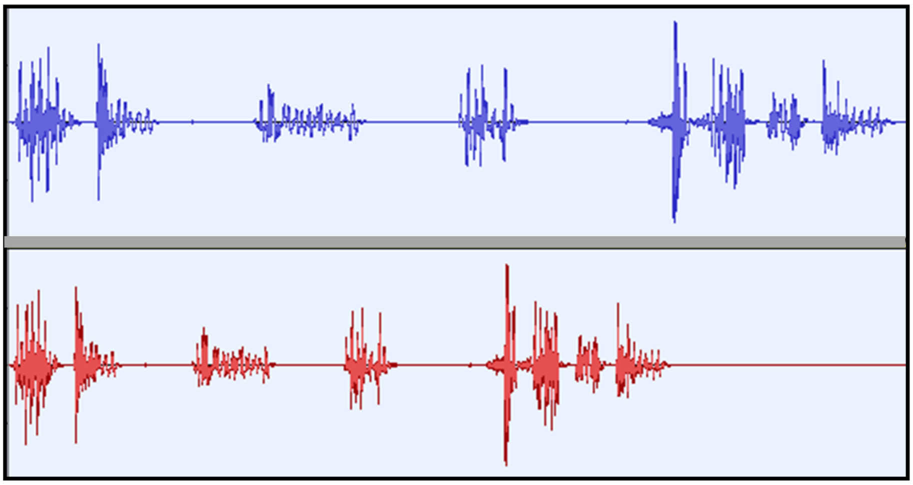


Figure 9.3: *Example of audio augmentation in the form of speed change.*

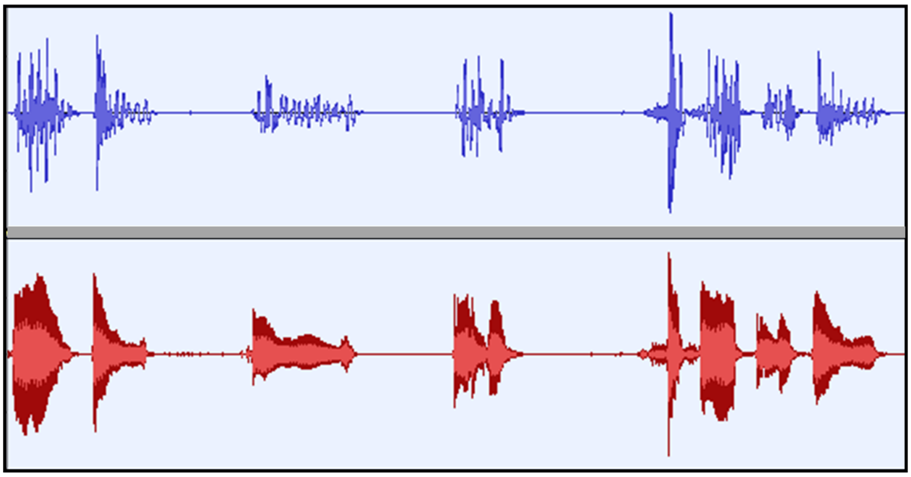


Figure 9.4: *Example of audio augmentation in the form of pitch change.*

source of audio recordings was 3638 phone conversations. Each call was five minutes long. The total recorded speech was 300 hours. The recruited participants were mostly college students from the United States. An automated operator connected two participants, and a suggested topic was given to start the discussion. There were 657 participants (46% male, 54% female). The sample rate for the Switchboard corpus was 6 KHz.

The four policies were based on three different audio manipulation techniques: time warp, time masking, and frequency masking. Table 9.13 outlines the four policies suggested by SpecAugment.

Figure 9.5 is the original mel spectrogram without any policies applied. Time warp was applied to the mel spectrogram, as shown in Figure 9.6. The mel spectrogram

Table 9.12: *Before augmentation: The distribution of the class labels in the new dataset. IR is the Imbalance Ratio score between the majority and each minority class.*

Trait/Label	M	H	L	IR M-H	IR M-L
Openness	46	11	11	4.18	4.18
conscientiousness	42	12	14	3.5	3
Extraversion	42	14	12	3	3.5
Agreeableness	45	10	13	4.5	3.46
Neuroticism	44	10	14	4.4	3.14

Table 9.13: *The four SpecAugment four policies.*

Policy	W	F	m_F	T	ρ	m_T
None	0	0	-	0	-	-
LB	80	27	1	100	1.0	1
LD	80	27	2	100	1.0	2
SM	40	15	2	70	0.2	2
SS	40	27	2	70	0.2	2

was wrapped to the left or right a distance of ω. ω is a distance chosen from a uniform distribution from 0 to ω.

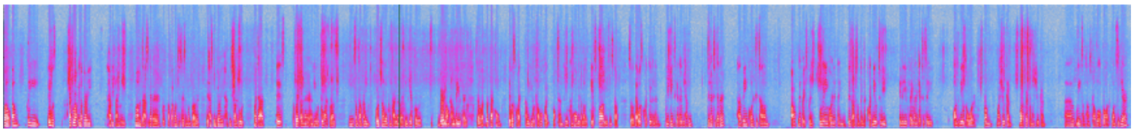


Figure 9.5: *Original mel spectrogram without any effects applied.*

Time masking was applied to the mel spectrogram (Figure 9.7). A number was chosen from a uniform distribution between 0 and τ. The selected time steps were masked. Another parameter was the number of time masks applied to each mel spectrogram.

Frequency masking uses the same technique as time masking. An f is chosen from a uniform distribution between 0 and f. Selected frequencies are masked from the mel spectrogram. An additional parameter for number of masks is also used. Figure 9.8 shows a single frequency mask applied to a mel spectrogram.

Figures 9.9 and 9.10 show a mel spectrogram after applying Librispeech Basic

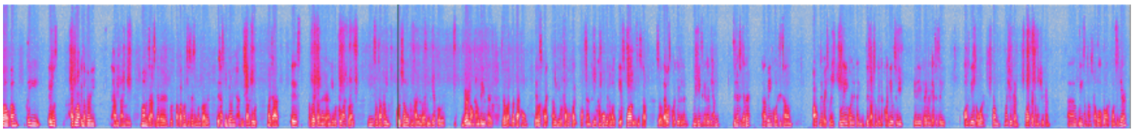


Figure 9.6: *Mel spectrogram after applying time warp effect.*

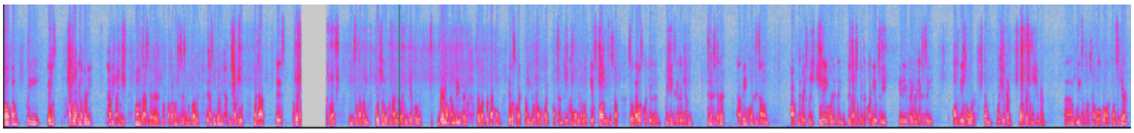


Figure 9.7: *Mel spectrogram after applying a single time mask.*

policy and Librispeech Double policy, respectively.

Two policies which were based on LibriSpeech were chosen to augment the new PTC: LibriSpeech Basic (LB) and LibriSpeech Double (LD). LibriSpeech policies are 16 KHz, which is the same sample rate as the PTC. The augmentation process was difficult. Figure 9.11 illustrates the process of augmentation. First, the wave file had to be converted to a mel spectrogram. Next, the mel spectrogram file was augmented twice with the Librispeech Basic policy. Then, the original mel spectrogram was augmented again with the Librispeech Double policy. Finally, both augmented mel spectrograms files were converted back to a WAV file format. This created more training data.

The minority labels were augmented using SpecAugment. For each trait, the minority classes (L and H) were doubled using LB policy and the doubled again with LD policy. The new IR score for the dataset after augmentation is shown in Table 9.14.

Table 9.14: *After augmentation: The distribution of the class labels in the new dataset. IR is the Imbalance Ratio score between the majority and each minority class.*

Trait/Label	**M**	**Aug H**	**Aug L**	**IR M-AH**	**IR M-AL**
Openness	46	31	31	1.35	1.48
Conscientiousness	42	32	34	1.31	1.23
Extraversion	42	34	32	1.23	1.31
Agreeableness	45	30	33	1.5	1.36
Neuroticism	44	30	34	1.46	1.29

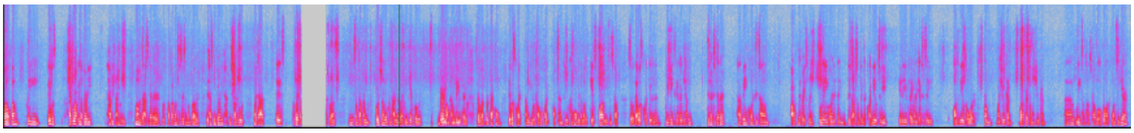


Figure 9.8: *Mel spectrogram after applying a single frequency mask.*

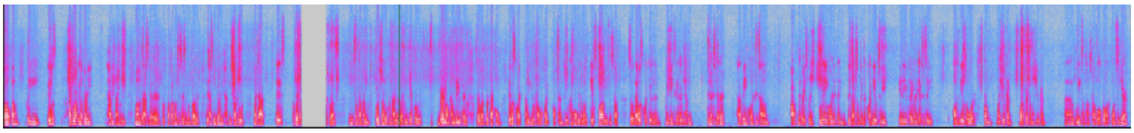


Figure 9.9: *Mel spectrogram after applying time warp, a single frequency mask and a single time mask.*

9.3.2 Augmented Personality Traits Corpus

Personality recognition is still a new field, and therefore multiple classifiers must be tested on this type of problem to choose the best-performing classifier or classifiers. The aim of the study was to find the top-performing classifier(s) for the PTC.

Therefore, eighteen classifiers were trained and tested on the new corpus. The different classifiers included the following: DL, SGD, LR, RF, Adaboost, SVM, SVC-linear, κNN, bagging, passive aggressive, ridge, ANN, perceptron, GB, LDA, QDA, naïve Bayes, and decision trees.

The classifiers, feature reduction techniques, evaluation measures, and significance were briefly explained in Chapter five.

All classifiers were hyperparameter tuned to produce the best possible recall. The corpus was divided as follows: 30% for testing on unseen data, and the remaining 70% was split again into validation (30%) and training (70%) data. Figure 9.12 explains how the dataset was divided and how the model was trained and built.

The 30% of the original corpus remained unseen by training or validation throughout all experiments. The augmentation was performed only on training and validation data. This was a multi-class problem, and thus the recall and AUC baselines were 33% and 50%, respectively.

The openness trait prediction results are shown in Table 9.15. The gradual increase in the amount of training data decreased the number of possible classifiers needed to predict openness trait from eighteen to only five. These classifiers were perceptron, deep leaning, κNN, linear SVC, and QDA. All classifiers except for κNN fluctuated between slightly higher or lower than 50% of UAR. The ρ-values and MCC

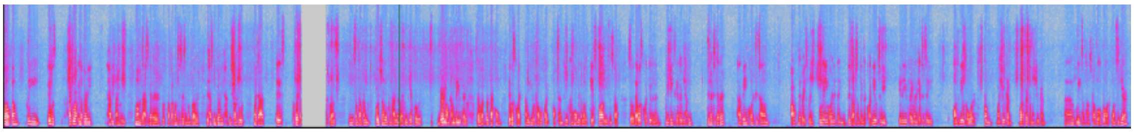


Figure 9.10: *Mel spectrogram after applying time warp, two frequency masks, and two time masks.*

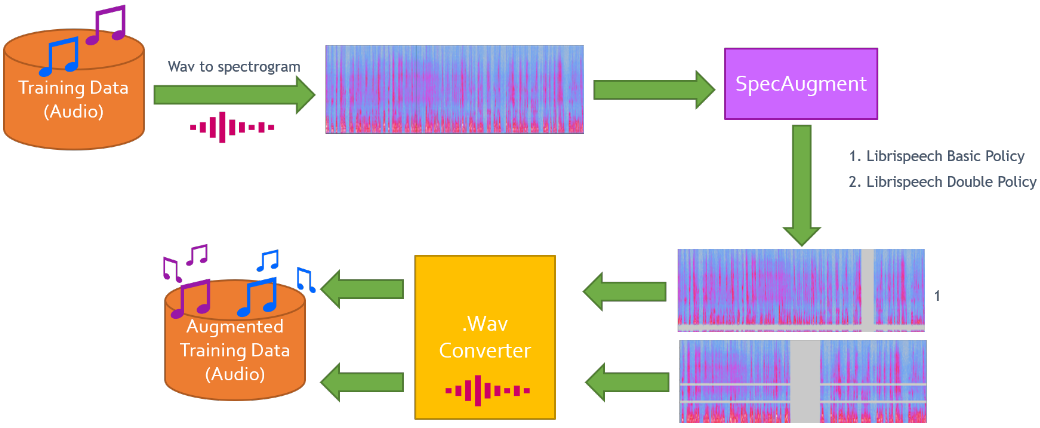

Training Data (Audio)
Wav to spectrogram
SpecAugment
1. Librispeech Basic Policy
2. Librispeech Double Policy
1
.Wav Converter
Augmented Training Data (Audio)

Figure 9.11: *The process of converting the WAV file to mel spectrogram, augmentation, and then convert back to a WAV file format.*

indicated instability with increased augmentation. This may have contributed to the classifiers' inability to deal with the complex nature and low separability of the data. κNN outperformed the other classifiers as the dataset was gradually augmented. The results became significant at 40% augmentation (ρ=0.002). MCC increased to 0.56, indicating that the classifier was predicting the correct class for most of the test set samples. Furthermore, the AUC was 72%, which was another indication of the classifier's ability to distinguish between the different class labels. At 60% augmentation, UAR jumps to 70% and AUC to 77%. The classifier significance was $\rho < 0.001$, and MCC was 0.63.

Table 9.16 displays the classifiers' performance with gradual augmentation for the conscientiousness trait. It is apparent from the results that the conscientiousness trait is hard to predict. Despite the augmentation, the best-performing classifiers were deep learning, κNN, Linear SVC, ridge, and QDA. The deep learning results were constant regardless of the augmentation percentage, as shown in the table. κNN performance improved as augmentation increased to 60% and then maintained its stability at 70% augmentation. κNN was also the top-performing classifier despite its mediocre UAR of 55%, with a ρ-value and MCC of 0.01 and 0.36, respectively. κNN's AUC of 66% supported by the MCC and ρ-value indicates the classifier's ability

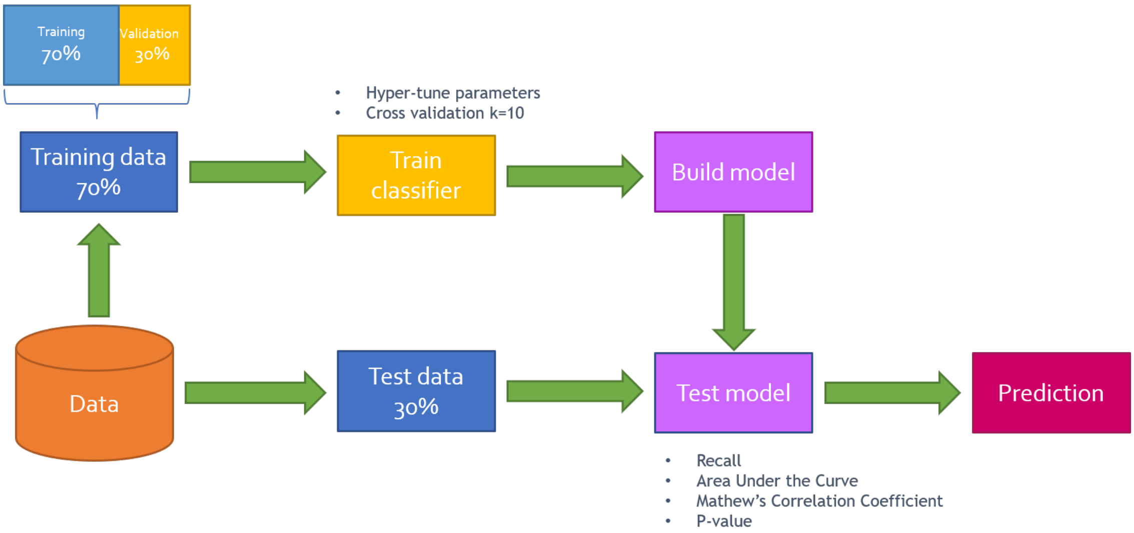


Figure 9.12: *The proposed framework model for personality recognition.*

to predict personality but with low performance. The remaining three classifiers fluctuated and did not produce any significant results regardless of augmentation.

Extraversion produced modest results with only three top classifiers, all of which are presented in Table 9.17. Two classifiers, perceptron and ANN, produced alternating high and low results as augmentation increased. The UAR of both was less than 50%, and the classifiers' uncertainty was reflected in their MCC and ρ-values. κNN maintained stability after 60% augmentation. Its UAR was 59% ($\rho - value = 0.002$). This performance was reinforced by the classifier's stability and ability to predict the actual class (MCC = 0.50, AUC = 69%).

Table 9.18 presents the results of the eighteen classifiers for the agreeableness trait. Preceptron, gradient boosting, and QDA produced results slightly higher than the baseline; however, they were unstable throughout the gradual augmentation. In contrast, κNN successfully predicted the agreeableness trait, with a UAR of 67%, and it was evident that the classifier could also predict the actual class (MCC = 0.55, $\rho - value = 0.0005$). The classifier was not predicting randomly and this was supported with its AUC at 74%.

For the neuroticism trait, similar to previous results, four out of five top-performing classifiers had results higher than baseline but not supported by MCC or AUC scores (Table 9.19). The results produced by the linear SVC, passive aggressive, Adaboost, and SGD classifiers fluctuated slightly higher than the baseline. The best-performing classifier was κNN (UAR = 67%, MCC = 0.53, $\rho - value < 0.001$). The classifier was not a random classifier based on its AUC score of 75%.

Table 9.15: *Openness trait prediction from gradually augmented dataset. Top-performing classifiers.*

	No Augment				10%				20%				30%				40%				50%				60%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	30.95%	48.12%	-0.13395	0.696404549	30.95%	48.39%	-0.05789474	0.69640455	34.29%	50.75%	0.03157895	0.56502	29.52%	46.98%	-0.1002	0.8073154	32.86%	49.48%	-0.0195	0.6964	34.29%	50.75%	0.03158	0.56502	34.29%	50.75%	0.03158	0.56502
Perceptron	43.33%	57.36%	0.145699	0.427061712	39.52%	54.48%	0.11336258	0.42706171	46.19%	59.48%	0.22373999	0.19127	49.52%	61.85%	0.27731	0.111943	50.48%	62.46%	0.27248	0.11194	42.86%	56.85%	0.17138	0.2983	50.95%	63.12%	0.36369	0.0595
ANN	33.33%	50.14%	0.01786	0.565017799	33.33%	50.28%	0.02564946	0.5650178	43.33%	57.50%	0.3	0.19127	43.33%	57.50%	0.3	0.1912718	40.00%	55.00%	0.24179	0.2983	40.00%	55.00%	0.24179	0.2983	40.00%	55.00%	0.24179	0.2983
Deep Learning	33.33%	50.00%	0	0.807315441	43.33%	57.50%	0.3	0.19127177	36.67%	52.50%	0.16966991	0.42706	40.00%	55.00%	0.24179	0.2983027	36.67%	52.50%	0.16967	0.42706	46.67%	60.00%	0.35012	0.11194	44.29%	58.12%	0.25465	0.19127
SVM	33.33%	50.00%	None	0.565017799	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502
Naïve Bayes	33.33%	50.14%	0.01786	0.565017799	28.10%	45.85%	-0.10859729	0.88998195	29.05%	46.33%	-0.0978741	0.88998	29.05%	46.33%	-0.0979	0.8899819	29.05%	46.33%	-0.0979	0.88998	29.05%	46.33%	-0.0979	0.88998	29.05%	46.33%	-0.0979	0.88998
Logistic Regression	36.67%	52.50%	0.16967	0.427061712	46.67%	32.56%	0.60138889	0.11194296	40.00%	55.14%	0.21052632	0.2983	40.00%	55.00%	0.24179	0.2983027	40.00%	55.00%	0.24179	0.2983	40.00%	55.00%	0.24179	0.2983	40.00%	55.00%	0.24179	0.2983
κNN	33.33%	50.00%	None	0.565017799	43.33%	57.50%	0.3	0.19127177	50.00%	62.50%	0.39659144	0.0595	50.00%	62.50%	0.39659	0.0594978	63.33%	72.50%	0.56043	0.00165	66.67%	75.00%	0.59874	0.0005	70.00%	77.50%	0.63733	0.00013
Bagging	33.33%	50.00%	None	0.565017799	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502
Random Forest	33.33%	50.00%	None	0.565017799	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502
Adaboost	33.33%	50.00%	None	0.565017799	33.33%	50.00%	None	0.5650178	34.29%	50.75%	0.03157895	0.56502	34.29%	50.62%	0.02565	0.5650178	30.95%	48.12%	-0.1339	0.6964	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502
Linear SVC	36.67%	52.50%	0.050809	0.696404549	44.29%	58.25%	0.24360849	0.19127177	45.24%	58.87%	0.22964606	0.19127	47.62%	60.62%	0.30934	0.111943	37.14%	52.46%	0.05496	0.56502	47.62%	60.62%	0.30934	0.11194	44.29%	58.12%	0.25465	0.19127
Passive Aggressive	28.57%	46.65%	-0.06275	0.944108823	47.62%	60.89%	0.29052007	0.11194296	39.52%	54.48%	0.11310925	0.42706	50.00%	62.50%	0.39659	0.0594978	44.29%	58.12%	0.25465	0.19127	43.33%	57.50%	0.3	0.19127	43.33%	57.50%	0.3	0.19127
Ridge	33.33%	50.00%	None	0.565017799	36.67%	52.50%	0.16966991	0.42706171	36.67%	52.50%	0.16966991	0.42706	36.67%	52.50%	0.16967	0.4270617	36.67%	52.50%	0.16967	0.42706	36.67%	52.50%	0.16967	0.42706	36.67%	52.50%	0.16967	0.42706
Gradient Boosting	33.33%	50.00%	None	0.565017799	33.33%	50.00%	None	0.5650178	30.95%	48.12%	-0.1339499	0.6964	34.29%	50.62%	0.02565	0.5650178	36.67%	52.64%	0.13362	0.42706	33.33%	50.00%	None	0.56502	33.33%	50.14%	0.01786	0.56502
LDA	26.67%	45.28%	-0.18086	0.990333346	36.67%	52.50%	0.16966991	0.42706171	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.5650178	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502	33.33%	50.00%	None	0.56502
QDA					41.90%	56.23%	0.12741453	0.5650178	59.05%	69.25%	0.38220419	0.02858	51.90%	63.45%	0.26226	0.1912718	57.62%	67.98%	0.35762	0.0595	55.24%	65.95%	0.3231	0.11194	55.71%	66.88%	0.3445	0.0595
SGD	34.29%	50.62%	0.010625	0.807315441	45.24%	58.73%	0.23519253	0.19127177	43.81%	57.60%	0.17076663	0.2983	45.24%	58.73%	0.23637	0.1912718	50.48%	62.46%	0.27294	0.11194	44.29%	58.12%	0.25636	0.19127	40.00%	55.00%	0.24179	0.2983

Table 9.16: *Conscientiousness trait prediction from gradually augmented dataset. Top-performing classifiers.*

	No Augment				10%				20%				30%				40%				50%				60%				70%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	39.32%	54.59%	0.12514493	0.29755609	28.67%	46.51%	-0.09901314	0.805908	33.55%	50.17%	0.007262	0.5624731	36.58%	52.38%	0.09586	0.42510874	33.55%	50.17%	0.00726	0.56247	40.07%	54.91%	0.13469	0.29756	37.04%	52.70%	0.074259852	0.42511	37.30%	53.08%	0.08076402	0.42510874
Perceptron	24.94%	43.72%	-0.13273624	0.94442013	48.87%	61.60%	0.259541766	0.061213	43.49%	57.50%	0.161431	0.1918168	50.76%	63.12%	0.35364	0.03009273	36.32%	52.28%	0.08515	0.42511	36.11%	52.14%	0.11098	0.42511	38.89%	54.23%	0.216126805	0.29756	36.58%	52.40%	0.09825497	0.42510874
ANN	33.80%	50.35%	0.01452329	0.56247314	36.32%	52.28%	0.09585851	0.425109	36.11%	52.11%	0.150626	0.4251087	36.11%	52.11%	0.15063	0.42510874	36.11%	52.11%	0.15063	0.42511	38.89%	54.23%	0.21613	0.29756	33.33%	50.06%	0.00814193	0.56247	36.11%	52.11%	0.1506257	0.42510874
Deep Learning	36.11%	52.20%	0.09585851	0.42510874	39.36%	54.55%	0.149949673	0.297556	38.89%	54.26%	0.181541	0.2975561	38.89%	54.23%	0.21613	0.29755609	38.89%	54.23%	0.21613	0.29756	38.89%	54.23%	0.21613	0.29756	38.89%	54.23%	0.216126805	0.29756	38.89%	54.23%	0.2161268	0.29755609
SVM	33.33%	50.00%	None	0.56247314	30.77%	48.05%	-0.1384128	0.694114	36.11%	52.11%	0.150626	0.4251087	36.11%	52.11%	0.15063	0.42510874	36.11%	52.11%	0.15063	0.42511	36.11%	52.11%	0.15063	0.42511	36.11%	52.11%	0.150625698	0.42511	36.11%	52.11%	0.1506257	0.42510874
Naïve Bayes	33.33%	50.00%	None	0.56247314	36.11%	52.11%	0.150625698	0.425109	38.89%	54.23%	0.216127	0.2975561	38.89%	54.23%	0.21613	0.29755609	38.89%	54.23%	0.21613	0.29756	38.89%	54.23%	0.21613	0.29756	36.11%	52.11%	0.150625698	0.42511	33.55%	50.14%	0.00584126	0.56247314
Logistic Regression	33.33%	50.00%	None	0.56247314	33.33%	50.00%	None	0.562473	36.11%	52.11%	0.150626	0.4251087	36.11%	52.11%	0.15063	0.42510874	38.89%	54.23%	0.21613	0.29756	36.11%	52.11%	0.15063	0.42511	36.11%	52.11%	0.150625698	0.42511	36.11%	52.11%	0.1506257	0.42510874
κNN	28.67%	46.51%	-0.1077076	0.80590751	34.73%	50.93%	0.023665266	0.562473	37.76%	53.17%	0.083678	0.4251087	40.79%	55.41%	0.13905	0.29755609	43.82%	57.65%	0.19119	0.19182	46.60%	59.71%	0.23936	0.11337	55.19%	66.17%	0.365145636	0.01343	55.19%	66.17%	0.36514564	0.01342875
Bagging	33.33%	50.00%	None	0.56247314	33.33%	50.93%	0.023665266	0.562473	33.33%	50.00%	None	0.5624731	33.33%	50.00%	None	0.56247314	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247314
Random Forest	33.33%	50.00%	None	0.56247314	33.33%	50.93%	0.023665266	0.562473	33.33%	50.00%	None	0.5624731	33.33%	50.00%	None	0.56247314	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247314
Adaboost	36.11%	52.17%	0.11307626	0.42510874	36.36%	52.24%	0.154696662	0.425109	33.33%	50.00%	None	0.5624731	33.33%	50.00%	None	0.56247314	36.11%	52.11%	0.15063	0.42511	30.77%	48.08%	-0.0957	0.69411	36.58%	52.40%	0.098254973	0.42511	33.33%	50.00%	None	0.56247314
Linear SVC	27.97%	45.93%	-0.08877747	0.88958725	48.41%	61.34%	0.270294694	0.061213	56.06%	67.22%	0.434909	0.0054228	50.76%	63.12%	0.35364	0.03009273	41.92%	56.50%	0.23778	0.19182	38.89%	54.23%	0.21613	0.29756	36.79%	52.54%	0.081618176	0.42511	39.36%	54.52%	0.163508	0.29755609
Passive Aggressive	30.03%	47.54%	-0.05083152	0.80590751	45.84%	59.39%	0.216757373	0.113374	50.51%	63.02%	0.341028	0.0300927	44.70%	58.64%	0.26738	0.11337432	38.89%	54.26%	0.18154	0.29756	38.89%	54.23%	0.21613	0.29756	39.36%	54.49%	0.162026648	0.29756	39.14%	54.35%	0.21745434	0.29755609
Ridge	33.33%	50.00%	None	0.56247314	36.36%	52.24%	0.154696662	0.425109	36.11%	52.11%	0.150626	0.4251087	38.89%	54.23%	0.21613	0.29755609	36.11%	52.11%	0.15063	0.42511	36.11%	52.11%	0.15063	0.42511	41.92%	56.47%	0.268403828	0.19182	41.92%	56.47%	0.26840383	0.19181679
Gradient Boosting	33.33%	50.00%	None	0.56247314	33.33%	50.93%	0.023665266	0.562473	33.33%	50.00%	None	0.5624731	33.33%	50.00%	None	0.56247314	33.33%	50.00%	None	0.56247	33.33%	50.03%	0.00407	0.56247	30.77%	48.05%	-0.138412803	0.69411	33.33%	50.03%	0.00407096	0.56247314
LDA	33.33%	50.03%	0.00407096	0.56247314	33.33%	50.93%	0.023665266	0.562473	33.33%	50.00%	None	0.5624731	33.33%	50.00%	None	0.56247314	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247314
QDA					24.61%	43.59%	-0.12514493	0.94442	33.20%	49.92%	-0.001185	0.694114	39.34%	54.57%	0.09586	0.42510874	44.21%	58.00%	0.16156	0.19182	45.82%	59.18%	0.18642	0.19182	55.32%	66.39%	0.333014132	0.01343	47.49%	60.62%	0.21740101	0.11337432
SGD	27.97%	45.93%	-0.08877747	0.88958725	49.13%	61.81%	0.25382854	0.061213	48.15%	61.18%	0.26623	0.0612131	37.04%	52.70%	0.07426	0.42510874	44.95%	58.74%	0.28661	0.11337	36.11%	52.11%	0.15063	0.42511	39.36%	54.52%	0.163508	0.29756	38.89%	54.23%	0.2161268	0.29755609

Table 9.17: *Extraversion trait prediction from gradually augmented dataset. Top-performing classifiers.*

	No Augment				10%				20%				30%				40%				50%				60%				70%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	28.21%	46.11%	-0.14476	0.805908	35.41%	51.30%	0.024905	0.562473	36.83%	52.53%	0.10166	0.42511	36.83%	52.50%	0.09825	0.42511	36.83%	52.50%	0.09825	0.42511	33.80%	50.29%	0.00959	0.56247	33.80%	50.29%	0.00959	0.56247	30.77%	48.08%	-0.0799	0.69411
Perceptron	41.61%	56.41%	0.13543	0.297556	37.47%	53.09%	0.069124	0.425109	40.25%	55.03%	0.12177	0.29756	43.32%	57.28%	0.18118	0.19182	46.35%	59.55%	0.23132	0.11337	49.38%	61.82%	0.2814	0.06121	49.13%	61.67%	0.27395	0.06121	46.13%	59.42%	0.25195	0.11337
ANN	25.64%	44.15%	-0.2469	0.889587	30.77%	48.05%	-0.13841	0.694114	36.36%	52.24%	0.1547	0.42511	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	36.36%	52.24%	0.1547	0.42511	33.33%	50.00%	None	0.56247	39.39%	54.48%	0.22197	0.29756
Deep Learning	34.48%	50.90%	0.022866	0.562473	39.61%	54.64%	0.166235	0.297556	36.83%	52.53%	0.10166	0.42511	33.80%	50.29%	0.01168	0.56247	31.24%	48.34%	-0.0726	0.69411	34.27%	50.58%	0.01703	0.56247	34.27%	50.58%	0.01703	0.56247	34.27%	50.58%	0.01703	0.56247
SVM	33.33%	50.00%	None	0.562473	30.77%	48.02%	-0.14248	0.694114	30.77%	48.02%	-0.1425	0.69411	30.77%	48.02%	-0.1425	0.69411	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247
Naïve Bayes	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247
Logistic Regression	33.33%	50.00%	None	0.562473	30.77%	48.02%	-0.14248	0.694114	30.77%	48.02%	-0.1425	0.69411	30.77%	48.02%	-0.1425	0.69411	30.77%	48.05%	-0.1384	0.69411	31.24%	48.34%	-0.0726	0.69411	33.80%	50.29%	0.01168	0.56247	33.80%	50.29%	0.01168	0.56247
κNN	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	36.11%	52.11%	0.15063	0.42511	44.70%	58.58%	0.31353	0.11337	50.51%	62.93%	0.39346	0.03009	56.57%	67.41%	0.46639	0.00542	59.34%	69.53%	0.50048	0.00198	59.34%	69.53%	0.50048	0.00198
Bagging	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247
Random Forest	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247
Adaboost	28.21%	46.10%	-0.1986	0.805908	28.42%	46.18%	-0.14476	0.805908	33.33%	50.00%	None	0.56247	36.36%	52.27%	0.11308	0.42511	36.36%	52.24%	0.1547	0.42511	36.36%	52.24%	0.1547	0.42511	36.36%	52.24%	0.1547	0.42511	33.33%	50.00%	None	0.56247
Linear SVC	41.61%	56.44%	0.136287	0.297556	34.69%	50.83%	0.016526	0.562473	42.60%	56.90%	0.18499	0.19182	39.57%	54.63%	0.13989	0.29756	43.32%	57.31%	0.18336	0.19182	40.54%	55.23%	0.13786	0.29756	31.70%	48.63%	-0.0503	0.69411	31.70%	48.63%	-0.0503	0.69411
Passive Aggressive	50.91%	63.15%	0.2708	0.061213	37.26%	52.84%	0.070609	0.425109	39.82%	54.75%	0.14052	0.29756	33.76%	50.27%	0.00852	0.56247	37.51%	52.96%	0.07921	0.42511	34.73%	50.87%	0.02154	0.56247	34.73%	50.87%	0.02154	0.56247	31.70%	48.63%	-0.0503	0.69411
Ridge	33.33%	50.00%	None	0.562473	30.77%	48.02%	-0.14248	0.694114	30.77%	48.02%	-0.1425	0.69411	30.77%	48.02%	-0.1425	0.69411	30.77%	48.05%	-0.1384	0.69411	28.21%	46.10%	-0.1986	0.80591	28.21%	46.10%	-0.1986	0.80591	37.30%	52.82%	0.08899	0.42511
Gradient Boosting	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	33.33%	50.12%	0.00959	0.56247	33.33%	50.03%	0.00407	0.56247	36.83%	52.50%	0.09825	0.42511	33.33%	50.03%	0.00407	0.56247	34.01%	50.43%	0.01146	0.56247	36.11%	52.14%	0.11098	0.42511
LDA	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.562473	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247	33.33%	50.00%	None	0.56247
QDA					36.23%	52.24%	0.04673	0.562473	40.17%	55.43%	0.12313	0.29756	37.70%	53.23%	0.0737	0.56247	41.10%	56.04%	0.12692	0.29756	40.50%	55.62%	0.12429	0.29756	61.81%	71.23%	0.42524	0.00198	37.65%	53.33%	0.07261	0.42511
SGD	44.64%	58.56%	0.175312	0.191817	37.47%	53.03%	0.069206	0.425109	40.71%	55.38%	0.11619	0.29756	48.15%	61.09%	0.28694	0.06121	37.04%	52.70%	0.08595	0.42511	37.30%	52.82%	0.08899	0.42511	40.07%	54.94%	0.14697	0.29756	37.30%	52.82%	0.08899	0.42511

Table 9.18: *Agreeableness trait prediction from gradually augmented dataset. Top-performing classifiers.*

	No Augment				10%				20%				30%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	33.73%	50.41%	0.017916	0.56386	46.30%	59.66%	0.339448	0.114313	35.05%	51.46%	0.04496	0.56386	38.76%	54.12%	0.11756	0.42725
Perceptron	33.47%	49.44%	-0.02497	0.887746	43.25%	56.99%	0.157899	0.299739	40.61%	55.45%	0.16267	0.29974	34.52%	50.80%	0.02269	0.56386
ANN	33.33%	50.00%	None	0.56386	33.73%	50.35%	0.021326	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Deep Learning	41.67%	56.70%	0.256692	0.193483	37.83%	53.50%	0.109062	0.427246	40.21%	55.10%	0.17924	0.29974	37.43%	53.00%	0.1184	0.42725
SVM	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Naïve Bayes	33.33%	50.00%	None	0.56386	33.73%	50.27%	0.012278	0.56386	31.35%	48.35%	-0.0738	0.69427	31.35%	48.35%	-0.0738	0.69427
Logistic Regression	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	30.95%	48.08%	-0.1411	0.69427	30.95%	48.17%	-0.1283	0.69427
κNN	30.95%	48.32%	-0.07674	0.694265	36.51%	52.53%	0.11084	0.427246	43.92%	57.82%	0.24758	0.19348	54.10%	65.30%	0.39543	0.03005
Bagging	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Random Forest	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Adaboost	33.33%	50.00%	None	0.56386	37.04%	52.80%	0.138133	0.427246	42.06%	56.88%	0.22119	0.19348	27.51%	45.78%	-0.1103	0.88775
Linear SVC	35.85%	51.36%	0.016129	0.804825	41.53%	55.99%	0.17376	0.299739	45.77%	59.54%	0.29206	0.11431	46.69%	60.01%	0.2968	0.11431
Passive Aggressive	30.95%	48.37%	-0.03	0.887746	44.31%	58.18%	0.226515	0.193483	42.06%	56.82%	0.23416	0.19348	47.09%	60.27%	0.27073	0.11431
Ridge	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Gradient Boosting	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	30.95%	48.24%	-0.0853	0.69427
LDA	34.66%	50.88%	0.030229	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
QDA					31.48%	49.18%	-0.005	0.804825	39.95%	54.66%	0.09014	0.69427	47.35%	59.73%	0.1844	0.42725
SGD	41.01%	55.37%	0.099751	0.56386	34.39%	50.33%	-0.00313	0.694265	40.61%	55.36%	0.15843	0.29974	47.49%	60.54%	0.25601	0.11431

	40%				50%				60%				70%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	37.83%	53.64%	0.10315	0.42725	34.13%	51.06%	0.04102	0.56386	32.54%	49.15%	-0.0277	0.69427	32.54%	49.15%	-0.027714	0.694265
Perceptron	46.16%	59.73%	0.26552	0.11431	38.89%	54.37%	0.22715	0.29974	39.29%	54.72%	0.17924	0.29974	42.06%	56.90%	0.2371257	0.193483
ANN	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Deep Learning	39.81%	54.83%	0.23459	0.29974	33.33%	50.00%	None	0.56386	30.95%	48.17%	-0.1283	0.69427	30.95%	48.17%	-0.128304	0.694265
SVM	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Naïve Bayes	30.95%	48.08%	-0.1411	0.69427	30.95%	48.08%	-0.1411	0.69427	28.57%	46.17%	-0.2026	0.80483	28.57%	46.17%	-0.202595	0.804825
Logistic Regression	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	30.95%	48.08%	-0.1411	0.69427	33.33%	50.00%	None	0.56386
κNN	60.58%	70.13%	0.47908	0.00526	64.29%	72.77%	0.52245	0.00187	67.06%	74.95%	0.55857	0.00059	67.06%	74.95%	0.5585652	0.000592
Bagging	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Random Forest	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Adaboost	39.29%	54.63%	0.17468	0.29974	33.33%	50.00%	None	0.56386	36.11%	52.18%	0.15824	0.42725	33.33%	50.00%	None	0.56386
Linear SVC	50.93%	63.56%	0.41939	0.03005	38.89%	54.37%	0.22715	0.29974	36.51%	52.53%	0.11084	0.42725	30.95%	48.17%	-0.128304	0.694265
Passive Aggressive	44.84%	59.00%	0.28736	0.11431	38.89%	54.37%	0.22715	0.29974	33.73%	50.35%	0.02133	0.56386	30.95%	48.17%	-0.128304	0.694265
Ridge	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
Gradient Boosting	42.59%	57.01%	0.28718	0.19348	42.59%	57.01%	0.28718	0.19348	38.89%	54.52%	0.19901	0.29974	42.59%	57.01%	0.2871764	0.193483
LDA	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386	33.33%	50.00%	None	0.56386
QDA	45.77%	59.15%	0.1799	0.29974	38.10%	52.97%	0.04874	0.80483	45.37%	58.68%	0.16931	0.42725	43.78%	57.62%	0.1510464	0.427246
SGD	40.21%	55.10%	0.17924	0.29974	41.67%	56.55%	0.28252	0.19348	36.51%	52.53%	0.11084	0.42725	36.51%	52.53%	0.11084	0.427246

Table 9.19: *Neuroticism trait prediction from gradually augmented dataset. Top-performing classifiers.*

	No Augment				10%				20%				30%				40%				50%				60%				70%			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	30.77%	47.96%	-0.1495	0.695898	30.77%	47.96%	-0.07828	0.695898	34.47%	50.80%	0.02652	0.56345	37.04%	52.73%	0.10819	0.42497	37.04%	52.73%	0.10819	0.42497	37.04%	52.61%	0.09812	0.42497	37.04%	52.61%	0.09812	0.42497	37.04%	52.61%	0.09812	0.42497
Perceptron	33.90%	49.70%	-0.02222	0.807973	43.02%	56.59%	0.1266	0.296389	51.85%	63.40%	0.25908	0.05955	51.85%	63.28%	0.273	0.05955	60.68%	69.85%	0.44893	0.00502	55.56%	65.77%	0.36009	0.02889	43.02%	56.48%	0.16755	0.29639	39.32%	53.87%	0.10236	0.42497
ANN	39.60%	54.77%	0.196245	0.296389	37.04%	52.73%	0.13183	0.424971	37.04%	52.61%	0.16659	0.42497	37.04%	52.61%	0.16659	0.42497	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345
Deep Learning	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	37.04%	52.61%	0.16659	0.42497	37.04%	52.61%	0.16659	0.42497	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	30.77%	47.96%	-0.1495	0.6959	33.33%	50.00%	None	0.56345
SVM	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345
Naïve Bayes	33.33%	50.00%	None	0.563445	30.77%	47.96%	-0.1495	0.695898	33.33%	50.00%	0	0.56345	33.33%	50.00%	0	0.56345	35.90%	52.04%	0.1495	0.42497	35.90%	52.04%	0.1495	0.42497	33.33%	50.00%	0	0.56345	35.90%	52.04%	0.1495	0.42497
Logistic Regression	33.33%	50.00%	None	0.563445	30.77%	48.08%	-0.13242	0.695898	40.74%	55.34%	0.18123	0.29639	38.18%	53.30%	0.11948	0.42497	40.74%	55.22%	0.23913	0.29639	33.33%	50.00%	None	0.56345	34.47%	50.69%	0.02453	0.56345	37.04%	52.61%	0.16659	0.42497
κNN	30.77%	47.96%	-0.08897	0.695898	31.91%	48.53%	-0.07799	0.695898	41.88%	55.79%	0.15098	0.29639	50.17%	62.48%	0.29849	0.05955	56.98%	67.13%	0.38494	0.01268	62.11%	71.21%	0.46212	0.00178	64.67%	73.25%	0.50052	0.00057	67.24%	75.28%	0.53919	0.00016
Bagging	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345
Random Forest	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345
Adaboost	33.33%	50.00%	None	0.563445	35.90%	52.04%	0.149502	0.424971	34.47%	50.69%	0.02453	0.56345	41.88%	55.91%	0.19236	0.29639	44.16%	57.51%	0.17328	0.29639	48.15%	60.44%	0.34892	0.11145	39.32%	53.98%	0.11185	0.42497	45.58%	58.51%	0.2542	0.19002
Linear SVC	33.90%	49.82%	-0.01735	0.807973	48.15%	60.44%	0.234729	0.111445	50.71%	62.59%	0.27541	0.05955	56.98%	67.24%	0.37931	0.01268	50.71%	62.48%	0.32985	0.05955	44.44%	57.83%	0.23331	0.19002	45.58%	58.51%	0.21692	0.19002	48.15%	60.44%	0.28801	0.11145
Passive Aggressive	37.61%	52.43%	0.03129	0.695898	48.15%	60.55%	0.250046	0.111445	48.15%	60.67%	0.23823	0.11145	50.71%	62.59%	0.29288	0.05955	49.57%	61.91%	0.3729	0.05955	50.71%	62.48%	0.32985	0.05955	39.32%	53.87%	0.10236	0.42497	48.15%	60.44%	0.34892	0.11145
Ridge	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	0	0.56345
Gradient Boosting	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.563445	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	35.90%	52.04%	0.08897	0.42497
LDA	33.33%	50.00%	None	0.563445	35.90%	52.04%	0.149502	0.424971	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345	33.33%	50.00%	None	0.56345
QDA					41.88%	56.49%	0.135793	0.296389	61.54%	70.34%	0.40097	0.01268	50.43%	62.39%	0.24827	0.11145	37.89%	53.56%	0.07833	0.56345	61.82%	70.89%	0.40745	0.00502	39.32%	54.10%	0.07619	0.42497	42.17%	57.50%	0.17104	0.19002
SGD	28.77%	45.74%	-0.10435	0.945745	50.71%	62.71%	0.289066	0.059555	51.85%	63.28%	0.273	0.05955	55.56%	66.01%	0.31514	0.02889	50.71%	62.83%	0.30713	0.05955	42.17%	56.69%	0.27424	0.19002	38.18%	53.30%	0.09372	0.42497	45.58%	58.40%	0.24285	0.19002

Due to the massive results tables, which included the classifiers and all of the measures; only a subset is presented. Full tables and graphs can be found in Appendix D.

Experiment 3 showed that the augmentation improved the training of the classifiers and their ability to predict the personality traits when tested with unseen data. Figures 9.13, 9.14, 9.15, 9.16, and 9.17 show the top-performing classifiers for each personality trait and how their UAR performs against the baseline. QDA was the only classifier that was unable to performed if there were not enough samples to train. This was found in the performance tables for all personality traits.

Across all traits, κNN had the ability to predict the personality traits with a significant UAR. This was also supported with high MCC and high AUC scores. Both measures confirmed the classifier's ability to predict the actual class, and the classifier behaved far away from any randomness. κNN achieved significance with augmentation, and its results did not fluctuate as with other classifiers. As the augmentation percentage increased, the κNN performance improved. In contrast, for all traits, κNN maintained stability at 60% augmentation, except for the neuroticism trait. When augmentation was 70%, κNN demonstrated improved results.

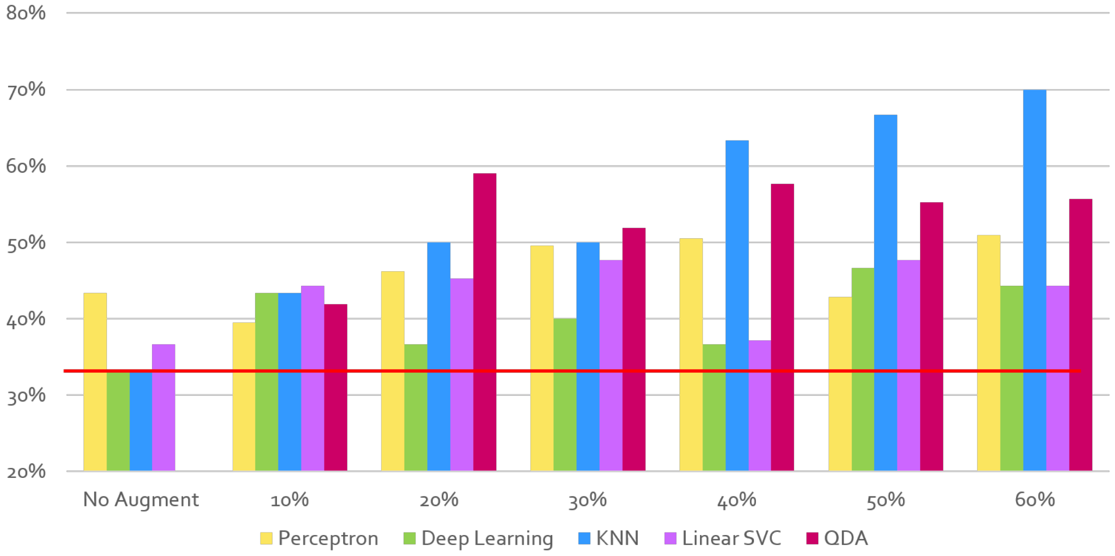


Figure 9.13: *Openness trait prediction with top classifiers.*

9.4 Experiment 4: Feature Reduction

This experiment was focused on selecting the best parameters to produce higher performance than naïve classification algorithms. Feature reduction techniques were

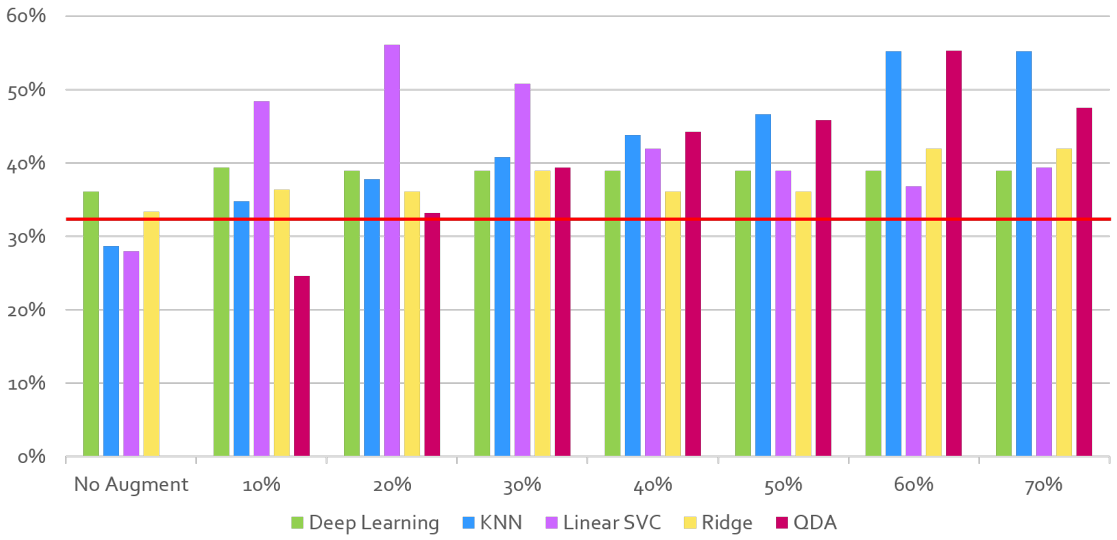


Figure 9.14: *Conscientiousness trait prediction with top classifiers.*

applied to study if performance would be improved.

Feature reduction techniques were applied to select the feature set that better enhanced the classifiers' performance. Feature reduction will either affect the performance positively or negatively. This experiment will add to the knowledge and understanding of how personality recognition and classifiers are affected by reduced feature sets.

Due to time limitations, ANOVA was chosen as a feature reduction technique. Feature reduction was performed on the highest augmented dataset for each personality trait.

Table 9.20 compares the prediction results of the openness trait when only selected features were used against the full feature set. Some classifiers, such as SVM, bagging, RF, Adaboost, and LDA failed to predict in both cases (selected feature set, full features) regardless. Four classifiers performed slightly better with selected features, such as decision tree and naïve Bayes. Perceptron was the best with selected features, with a UAR of 49%. The remaining ten classifiers performed better with all features.

As shown in Table 9.21, conscientiousness trait performance with all features was better than with only selected features. Only two classifiers performed slightly better with selected features: perceptron and SGD. In contrast, nine classifiers produced better UAR with all features included. In both feature scenarios, κNN had the highest UAR.

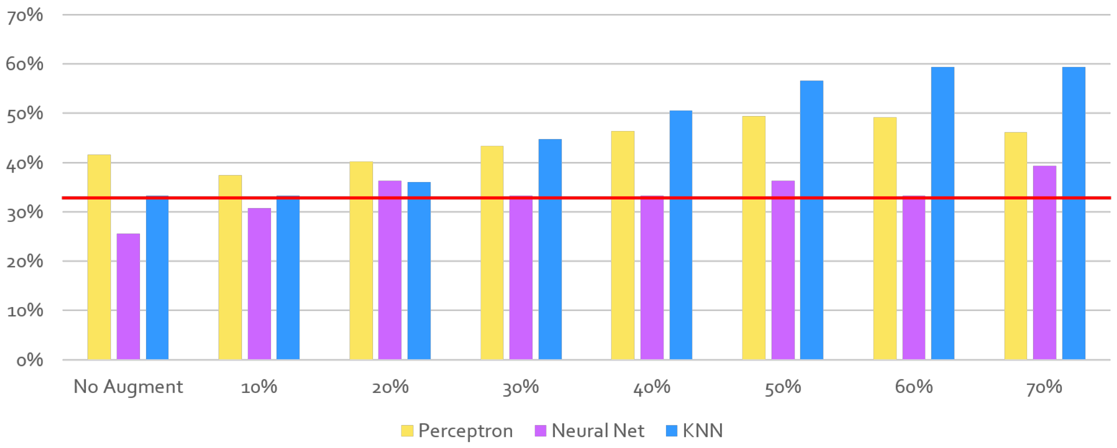


Figure 9.15: *Extraversion trait prediction with top classifiers.*

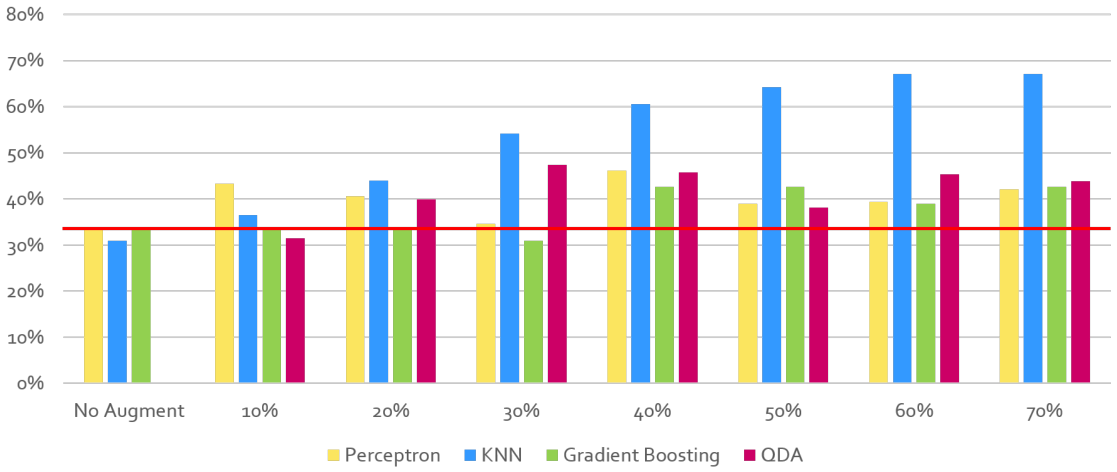


Figure 9.16: *Agreeableness trait prediction with top classifiers.*

A comparison of the extraversion prediction results between the use of selected features and all features is shown in Table 9.22. Only five classifiers performed slightly better with the selected features. Perceptron had the highest UAR with selected features (53%). In addition, only five classifiers performed better with all features included.

Table 9.23 shows that nine classifiers had slightly better UAR with the selected features. Moreover, only three classifiers produced higher UAR when all features were selected. SGD was the best classifier when there was a selected feature set. SGD had a UAR of 51%.

With selected features, only four classifiers exhibited slightly improved performance for the neuroticism trait (Table 9.24). Passive aggressive performed best with selected

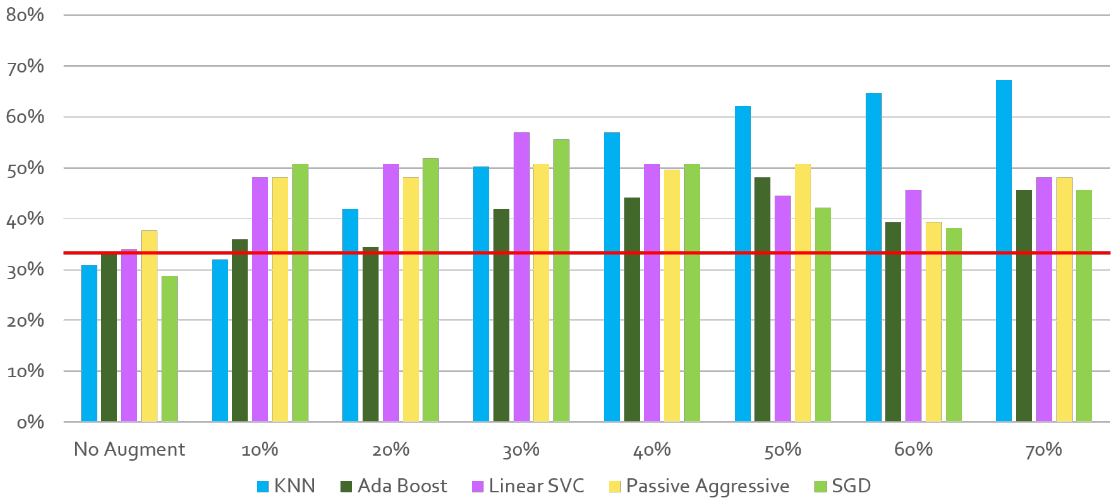


Figure 9.17: *Neuroticism trait prediction with top classifiers.*

features with UAR at 51%. Furthermore, the majority of the classifiers performed better when all features were included.

This experiment tested feature reduction techniques. The classifiers and the dataset were tested against two different scenarios. In scenario 1, selected features were fed into the classifiers, and in scenario 2, all features were inserted into the classifiers.

In the first scenario, selected features were chosen and used for classification. It was shown in the previous section that of the five personality traits no classifier had consistent performance. Classifiers were feature dependent and produced varying UAR scores, which could be a sign of instability.

The second scenario incorporated the full feature set. The κNN performed the best across all traits. This further proves the stability and power of κNN as a classifier for high-dimensional datasets with complex and low-separability datasets.

Figures 9.18, 9.19, 9.20, 9.21, and 9.22 show the difference in classifier performance using the UAR between incorporating selected features and all features. The best-performing classifier in both scenarios are listed in bold, and the cells are coloured for easier comparison.

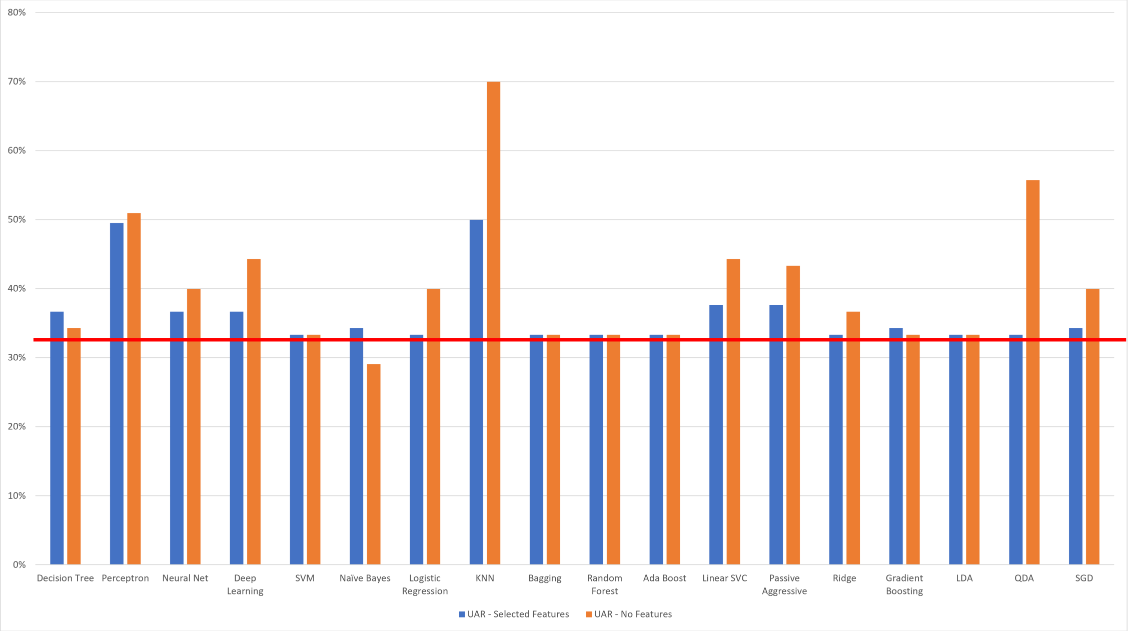


Figure 9.18: *Comparison of UAR openness trait prediction classifier performance with selected features and all features.*

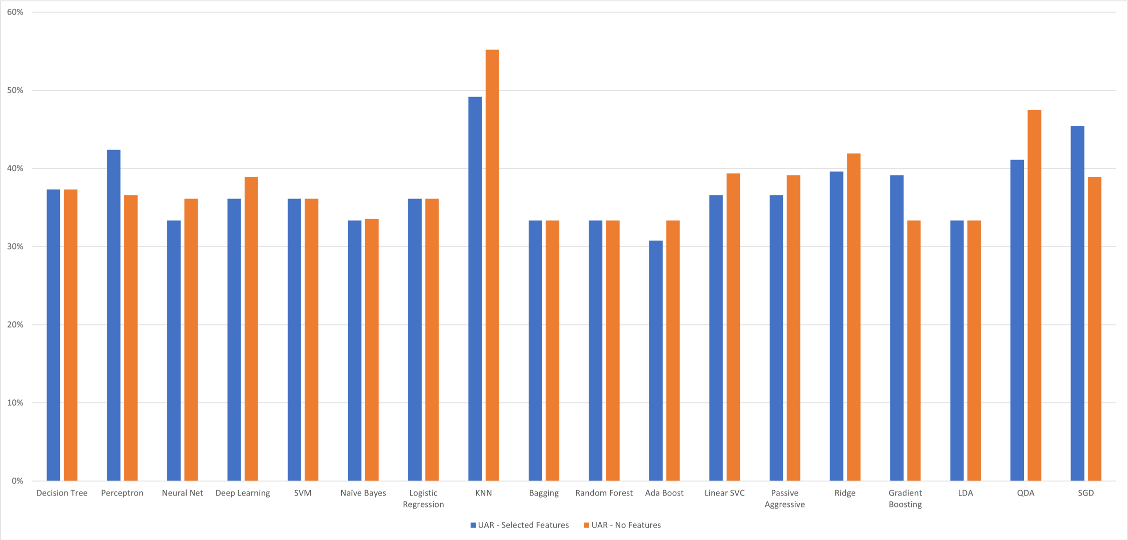


Figure 9.19: *Comparison of UAR conscientiousness trait prediction classifier performance with selected features and all features.*

Table 9.20: *Comparison of openness trait UAR on fully augmented dataset with selected features and all features.*

	Selected Features				All Features			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	36.67%	52.63%	0.1336	0.427	34.29%	50.75%	0.0315	0.565
Perceptron	49.52%	61.84%	0.2773	0.111	50.95%	63.12%	0.3636	0.059
ANN	36.67%	52.50%	0.1696	0.427	40.00%	55.00%	0.2417	0.298
Deep Learning	36.67%	52.50%	0.1696	0.427	44.29%	58.12%	0.2546	0.191
SVM	33.33%	50.00%	None	0.565	33.33%	50.00%	None	0.565
Naïve Bayes	34.29%	50.61%	0.0256	0.565	29.05%	46.33%	-0.0978	0.889
Logistic Regression	33.33%	50.00%	None	0.565	40.00%	55.00%	0.2417	0.298
κNN	**50.00%**	62.50%	0.3965	0.059	**70.00%**	77.50%	0.6373	0.0001
Bagging	33.33%	50.00%	None	0.565	33.33%	50.00%	None	0.565
Random Forest	33.33%	50.00%	None	0.565	33.33%	50.00%	None	0.565
Adaboost	33.33%	50.00%	None	0.565	33.33%	50.00%	None	0.565
Linear SVC	37.62%	53.11%	0.121	0.427	44.29%	58.12%	0.2546	0.191
Passive Aggressive	37.62%	53.11%	0.121	0.427	43.33%	57.50%	0.3	0.191
Ridge	33.33%	50.00%	None	0.565	36.67%	52.50%	0.1696	0.427
Gradient Boosting	34.29%	50.75%	0.0315	0.565	33.33%	50.14%	0.0178	0.565
LDA	33.33%	50.00%	None	0.565	33.33%	50.00%	None	0.565
QDA	33.33%	49.86%	-0.0058	0.807	55.71%	66.88%	0.3444	0.059
SGD	34.29%	50.61%	0.0256	0.560	40.00%	55.00%	0.2417	0.298

Table 9.21: *Comparison of conscientiousness trait UAR on fully augmented dataset with selected features and all features.*

	Selected Features				**All Features**			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	37.30%	53.08%	0.0807	0.4251	37.30%	53.08%	0.0808	0.4251
Perceptron	42.39%	56.75%	0.2201	0.1918	36.58%	52.40%	0.0983	0.4251
ANN	33.33%	50.00%	None	0.6941	36.11%	52.11%	0.1506	0.4251
Deep Learning	36.11%	52.11%	0.1506	0.4251	38.89%	54.23%	0.2161	0.2976
SVM	36.11%	52.11%	0.1506	0.4251	36.11%	52.11%	0.1506	0.4251
Naïve Bayes	33.33%	50.00%	None	0.5624	33.55%	50.14%	0.0058	0.5625
Logistic Regression	36.11%	52.11%	0.1506	0.4251	36.11%	52.11%	0.1506	0.4251
κNN	**49.16%**	61.71%	0.2922	0.0612	**55.19%**	66.17%	0.3651	0.0134
Bagging	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
Random Forest	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
Adaboost	30.77%	48.05%	-0.1384	0.69411	33.33%	50.00%	None	0.5625
Linear SVC	36.58%	52.40%	0.0982	0.4251	39.36%	54.52%	0.1635	0.2976
Passive Aggressive	36.58%	52.40%	0.0982	0.4251	39.14%	54.35%	0.2175	0.2976
Ridge	39.61%	54.64%	0.1662	0.2975	41.92%	56.47%	0.2684	0.1918
Gradient Boosting	39.14%	54.38%	0.1821	0.2975	33.33%	50.03%	0.0041	0.5625
LDA	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
QDA	41.10%	55.81%	0.1255	0.2975	47.49%	60.62%	0.2174	0.1134
SGD	45.42%	58.99%	0.2705	0.1133	38.89%	54.23%	0.2161	0.2976

Table 9.22: *Comparison of extraversion trait UAR on fully augmented dataset with selected features and all features.*

	Selected Features				All Features			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	30.77%	48.08%	-0.0798	0.6941	30.77%	48.08%	-0.0799	0.6941
Perceptron	**53.09%**	64.45%	0.3022	0.03	46.13%	59.42%	0.2520	0.1134
ANN	31.24%	48.31%	-0.0652	0.6941	39.39%	54.48%	0.2220	0.2976
Deep Learning	34.27%	50.58%	0.017	0.5624	34.27%	50.58%	0.0170	0.5625
SVM	30.77%	48.05%	-0.1384	0.6941	33.33%	50.00%	None	0.5625
Naïve Bayes	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
Logistic Regression	36.83%	52.53%	0.1016	0.4251	33.80%	50.29%	0.0117	0.5625
κNN	39.39%	54.47%	0.2219	0.2975	**59.34%**	69.53%	0.5005	0.0020
Bagging	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
Random Forest	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
Adaboost	39.39%	54.47%	0.2219	0.2975	33.33%	50.00%	None	0.5625
Linear SVC	34.73%	50.87%	0.0215	0.5624	31.70%	48.63%	-0.0503	0.6941
Passive Aggressive	34.73%	50.87%	0.0215	0.5624	31.70%	48.63%	-0.0503	0.6941
Ridge	37.04%	52.69%	0.0859	0.4251	37.30%	52.82%	0.0890	0.4251
Gradient Boosting	30.77%	48.08%	-0.0798	0.6941	36.11%	52.14%	0.1110	0.4251
LDA	33.33%	50.00%	None	0.5624	33.33%	50.00%	None	0.5625
QDA	30.71%	47.74%	-0.0538	0.8059	37.65%	53.33%	0.0726	0.4251
SGD	37.30%	52.82%	0.0889	0.4251	37.30%	52.82%	0.0890	0.4251

Table 9.23: *Comparison of agreeableness trait UAR on fully augmented dataset with selected features and all features.*

	Selected Features				**All Features**			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	42.99%	57.27%	0.2371	0.1934	32.54%	49.15%	-0.0277	0.6943
Perceptron	45.24%	58.63%	0.2328	0.1934	42.06%	56.90%	0.2371	0.1935
ANN	36.11%	52.18%	0.1582	0.4272	33.33%	50.00%	None	0.5639
Deep Learning	33.33%	50.00%	None	0.5638	30.95%	48.17%	-0.1283	0.6943
SVM	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
Naïve Bayes	33.33%	50.00%	None	0.5638	28.57%	46.17%	-0.2026	0.8048
Logistic Regression	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
κNN	42.59%	57.01%	0.2871	0.1934	**67.06%**	74.95%	0.5586	0.0006
Bagging	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
Random Forest	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
Adaboost	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
Linear SVC	36.11%	52.18%	0.1582	0.4272	30.95%	48.17%	-0.1283	0.6943
Passive Aggressive	33.33%	50.00%	None	0.5638	30.95%	48.17%	-0.1283	0.6943
Ridge	36.11%	52.18%	0.1582	0.4272	33.33%	50.00%	None	0.5639
Gradient Boosting	36.11%	52.18%	0.1582	0.4272	42.59%	57.01%	0.2872	0.1935
LDA	33.33%	50.00%	None	0.5638	33.33%	50.00%	None	0.5639
QDA	24.60%	43.26%	-0.1681	0.9424	43.78%	57.62%	0.1510	0.4272
SGD	**51.85%**	64.02%	0.4214	0.03	36.51%	52.53%	0.1108	0.4272

Table 9.24: *Comparison of neuroticism trait UAR on fully augmented dataset with selected features and all features.*

	Selected Features				**All Features**			
Classifier	**UAR**	**AUC**	**MCC**	**ρ-Value**	**UAR**	**AUC**	**MCC**	**ρ-Value**
Decision Tree	35.90%	52.15%	0.0861	0.4249	37.04%	52.61%	0.0981	0.4250
Perceptron	48.15%	60.43%	0.288	0.1114	39.32%	53.87%	0.1024	0.4250
ANN	33.33%	50.00%	None	0.5634	33.33%	50.00%	None	0.5634
Deep Learning	37.04%	52.60%	0.1665	0.4249	33.33%	50.00%	None	0.5634
SVM	33.33%	50.00%	None	0.5634	33.33%	50.00%	None	0.5634
Naïve Bayes	33.33%	50.00%	None	0.5634	35.90%	52.04%	0.1495	0.4250
Logistic Regression	33.33%	50.00%	None	0.5634	37.04%	52.61%	0.1666	0.4250
κNN	44.73%	58.72%	0.3182	0.1114	**67.24%**	75.28%	0.5392	0.0002
Bagging	33.33%	50.00%	None	0.5634	33.33%	50.00%	None	0.5634
Random Forest	33.33%	50.00%	None	0.5634	33.33%	50.00%	None	0.5634
Adaboost	43.02%	56.59%	0.1774	0.2963	45.58%	58.51%	0.2542	0.1900
Linear SVC	48.15%	60.43%	0.288	0.1114	48.15%	60.44%	0.2880	0.1114
Passive Aggressive	**51.85%**	63.04%	0.3383	0.0595	48.15%	60.44%	0.3489	0.1114
Ridge	44.44%	57.82%	0.2333	0.19	33.33%	50.00%	none	0.5634
Gradient Boosting	35.90%	52.03%	0.0889	0.4249	35.90%	52.04%	0.0890	0.4250
LDA	33.33%	50.00%	None	0.5634	33.33%	50.00%	None	0.5634
QDA	33.05%	50.03%	0.0061	0.6958	42.17%	57.50%	0.1710	0.1900
SGD	33.33%	50.00%	None	0.5634	45.58%	58.40%	0.2429	0.1900

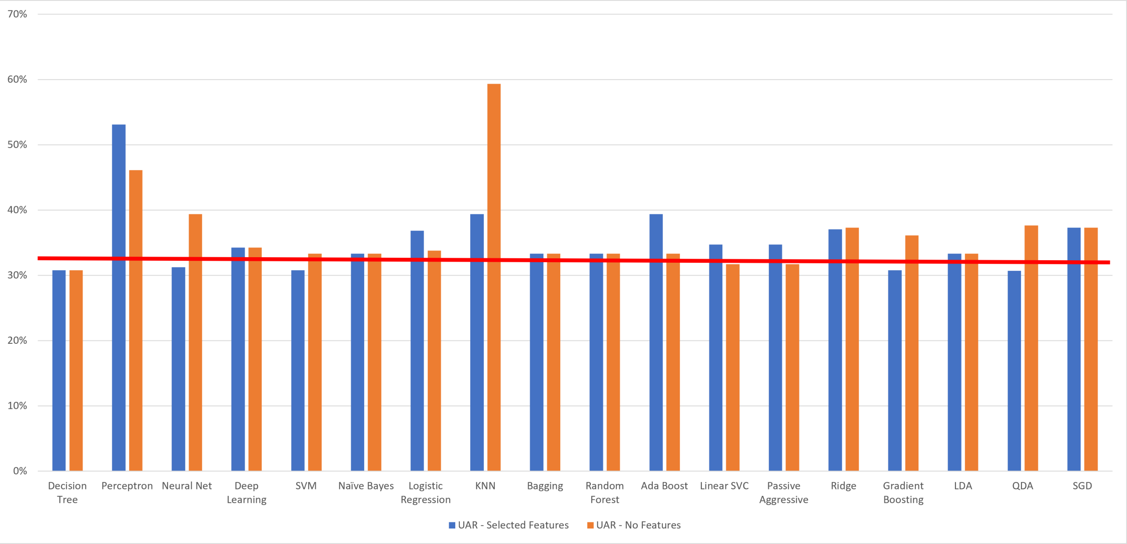


Figure 9.20: *Comparison of UAR extraversion trait prediction classifier performance with selected features and all features.*

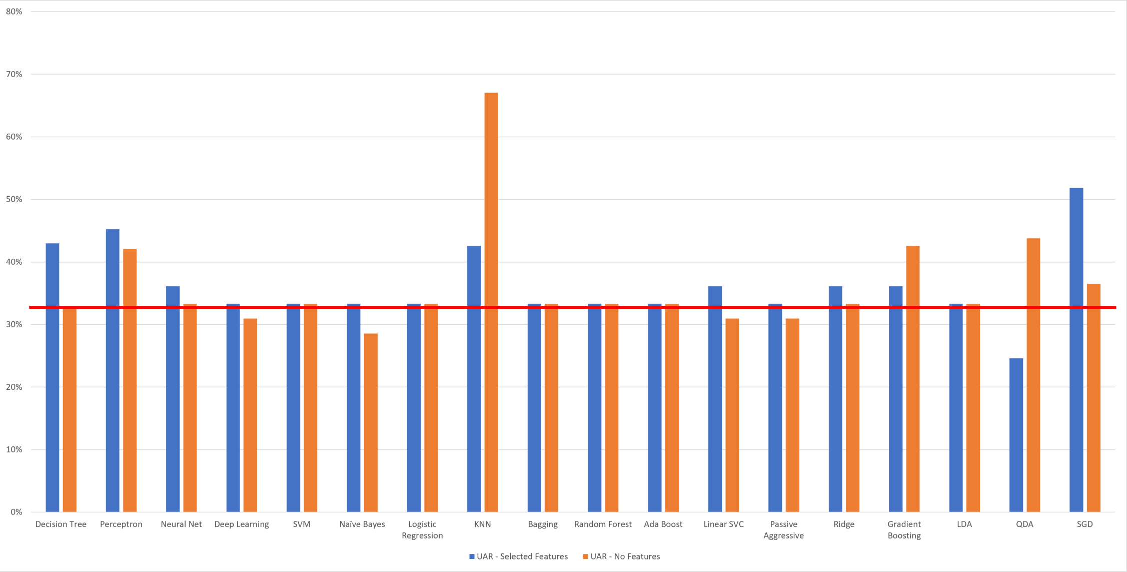


Figure 9.21: *Comparison of UAR agreeableness trait prediction classifier performance with selected features and all features.*

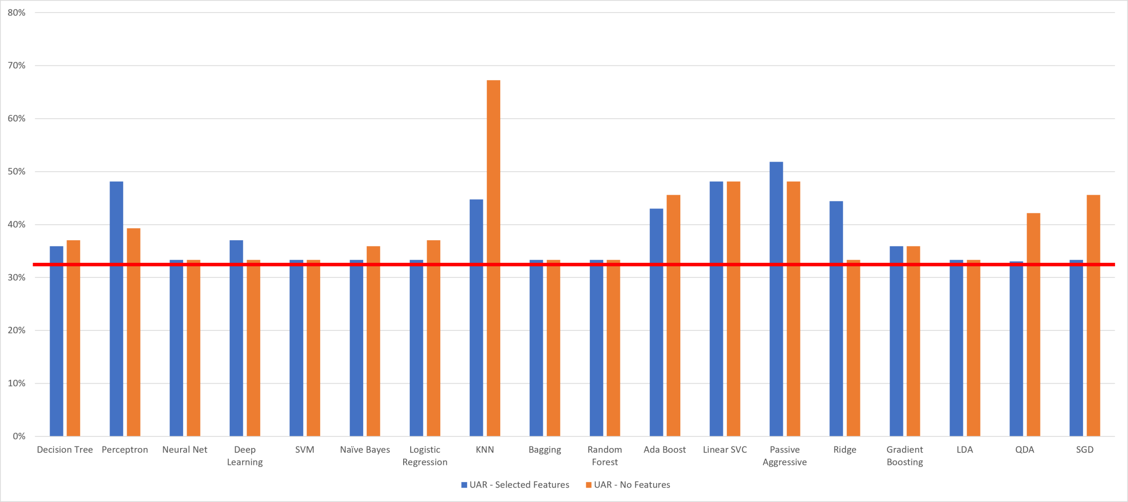


Figure 9.22: *Comparison of UAR neuroticism trait prediction classifier performance with selected features and all features.*

Chapter Summary

This chapter described four different experiments. Due to the unknown nature of the new corpus, eighteen classifiers were used for training, validating, and building the classification models.

For experiment 1, default setting, the results of the eighteen classifiers are shown in tables and figures for each personality trait. The results of experiment 1 were inconclusive due to imbalance and high dimensionality.

Experiment 2 adopted the holdout method. This method was used to avoid overfitting and the test set leaking during cross-validation. The test set was chosen and remained unseen during classifier training and building. The training set was split into training and validation sets. The training set was used to build the model and the validation set for hyperparameter tuning and enhancing performance. Hyperparameter tuning was applied for all eighteen classifiers across all five personality traits. The best parameters were chosen for all classifiers, and then they were trained and tested to yield improved UAR compared to experiment 1 (default setting).

Experiment 3 tackled the imbalance issue by applying data augmentation. The SpecAugment technique was used to create new training samples for minority classes. To understand the effect of data augmentation, a gradual test of data augmentation was adopted. Specifically, data were augmented at 10% intervals until maximum augmentation was achieved. It was evident that, as augmentation increased, κNN outperformed all of the other classifiers. Its UAR scores were significant, and its MCC score confirmed that the classifier was able to predict the actual class for the test sample. In addition, the AUC scores clearly indicated that κNN was not performing random predictions.

Finally, the aim of experiment 4 was to deal with high dimensionality and further improve the classifiers' prediction. Therefore, ANOVA was applied to select a feature subset and then train and test the augmented dataset. Some classifiers performed better with selected features than with the full feature set. Nevertheless, for selected features, no single classifier was the best across all personality traits. Furthermore, the results were not significant, and the MCC scores were low.

In contrast, with a full feature set, κNN remained stable across all personality traits as the best classifier. This was supported by very good MCC and AUC scores and significant ρ-values. Interestingly, κNN with the full feature set was the best-performing classifier with either selected features or a full feature set.

Chapter 10

Conclusion and Future Work

A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.

Alan Turing

The research described in this thesis began with a focus on personality recognition from honest signals. However, as the work progressed, several major issues in prior approaches emerged. The first was the unintentional misuse of several personality perception datasets for personality recognition. The second was the massive gap between the approaches to personality research in psychology and those in computer science; the understanding of personality was very different between the two communities. Therefore, it was necessary to first design and collect a new corpus before pursuing personality traits recognition and addressing the research questions proposed in the first chapter. This journey is described in the chapters of this research. Chapter One explained the initial goal, which was focused on identifying honest signals. Chapter Two explored several personality theories and identified several published articles on social signal research. Chapter Three highlighted the big five personality theory and presented several published articles on personality recognition. In Chapter Four, the thesis explained the current corpora for personality recognition. The experimental methodology was described in Chapter Five, which included the experimental setting, machine learning algorithms, feature reduction techniques, and evaluation metrics. Chapter Six explained how the experiment was conducted and the results. Chapter Seven presented a more sophisticated view of personality from psychology, including a background on personality research, recent psychological literature, and the RAM. Chapter Eight described the data collection protocol and presented the exploratory data analysis. Chapter Nine presented the new Personality Trait Corpus (PTC) and discussed the results of several experiments.

10.1 Research Findings

The research findings answer the research questions established at the beginning of this thesis.

- ***Research Question 1 (RQ1):*** **Can we build a corpus based on an accuracy model from psychology?** After researching the personality psychology community for theories and guidelines on accuracy, it was clear that the new PTC followed the guidelines for replicating data collection studies in the psychology community to ensure accurate personality recognition. This research adopted the Realistic Accuracy Model (RAM) to capture behaviour and produce an accurate personality judgment. So far, the new PTC is the only corpus that was built based on the RAM model. Furthermore, personality should be classified using at least three different class labels. Binary classification was not suitable for personality traits.
- ***Research Question 2 (RQ2):*** **Is automatic personality recognition through non-verbal acoustic features from speech possible?** The new PTC showed significant results, which performed well above the baseline. However, this was achieved when the corpus training data were augmented. It was evident that to answer this question with a greater degree of confidence will require the collection of more data. Automatic personality recognition is possible, but further data are required to increase credibility.
- ***Research Question 3 (RQ3):*** **Are there any specific features that can improve personality recognition from acoustic non-verbal cues from speech?** Feature reduction was performed on the new PTC using different feature reduction techniques. The results from a reduced feature set yielded non-significant results and rarely outperformed the baseline. It was found that when all features were included the classifiers produced significant results above the baseline.
- ***Research Question 4 (RQ4):*** **What is the best machine learning algorithm for personality recognition from non-verbal cues from speech?** The new PTC is a new dataset. Therefore, many classifiers were

tested to select the best-performing classifier. Evidently, some classifiers such as deep learning cannot be conclusive because they require a large training dataset. However, it was clearly shown that κNN significantly outperformed all classifiers. Apparently, κNN works best with more features to be able to distinguish the three-class labels. It also handled low separability and complex data well.

- ***Research Question 5 (RQ5):*** **What traits can be recognised accurately from non-verbal features from speech?** This research has shown that all big five personality traits can be recognised above baseline levels, and the results were significant. However, the trait that could be recognised most accurately was openness (UAR = 70%, $\rho - value < 0.0001$, MCC = 0.63). For agreeableness and neuroticism, UAR was 67%. The UAR for extraversion was at 59.34%, while conscientiousness performed worst, with a UAR of 55.19%. These results implied that openness is the easiest trait to predict from acoustic non-verbal cues, while conscientiousness is the hardest to predict.

- ***Research Question 6 (RQ6):*** **Is it possible to automatically detect non-verbal acoustic behavioural cues in data captured with sensors like microphones and headsets?** The new PTC appears capable of recognising personality. However, further research is required to collect more data and revisit the experimentation phase to obtain more credible and valid confirmation.

10.2 Novel Contributions

This research has produced potentially valuable contributions to both the computer science and personality psychology communities. These contributions are as follows:

- ***Experiments with the Speaker Personality Corpus (SPC):*** For experimental purposes, this research highlighted several limitations with the SPC and that it has been misused for personality recognition when it was created for the purpose of personality perception. The ground truth was for the perception of zero-acquaintance judges with no information about or relation to the target.

- ***Personality Traits Corpus:*** This research has collected and built a dataset for personality recognition based on theories, guidelines, and knowledge from the personality psychology research community. The new PTC has very rich data despite its small size. This research has produced a corpus that can be used for future quantitative and qualitative research.

- ***Classification models:*** An important contribution was building the classification models from the new corpus. Eighteen different models were built for each personality trait. The classifiers were hyperparameter tuned to yield the best performance. The κNN was the best-performing classifier, with a full feature set across all personality traits.

- ***Review previous research on personality recognition:*** This research has explored previous work on personality recognition. It has highlighted research related to the big five in both verbal and non-verbal models. It has also presented recent research in social signals. This thesis can serve as a reference for scientists interested in pursuing research on both personality recognition and social signals.

10.3 Future Work

Automatic personality recognition remains in its infancy. However, this research has aimed to bridge the gap between personality psychology and computer science. It has reduced the gap by building automatic personality recognition corpus and models based on guidelines and theories from personality psychology. This research has taken personality recognition a step further, and many future possibilities and opportunities are now possible. An attractive opportunity would be transcribing the corpus, building a model for personality recognition, and comparing the results with what this research achieved. Some interesting opportunities for research include expanding the corpus and adding more demographic data, such as education and employment data. This raises an interesting research question regarding the effect of personality and education. An intriguing future possibility is another study focused on the ‘moderators of accuracy’. There are four moderators of accuracy that aid in accurately judging a personality: good judge, good trait, good target, and good information. Each of these moderators affects the RAM at different stages. RAM and its moderators are a rich topic to investigate, which will extend computer science research on automatic personality recognition.

Moreover, it would be interesting if spectrogram images were extracted from audio files. This could open another window of research opportunities for personality recognition of acoustic features from images. This method of transforming audio to images could offer a new way to analyse acoustic features and recognise personality.

In addition, HRI offers an excellent a great opportunity to research and implement personality traits into robots. If robots can manifest personality based on the person they are interacting with, they could be of assistance to nursing homes, healthcare professionals in different fields, learning and education, and children with autism or other disorders.

10.4 Challenges

This section lists the challenges associated with the data collection and the limitations of this thesis. Many of the challenges were due to COVID-19. The logistics of data collection and the methods changed several times, from face-to-face methods to online-based interviews. Therefore, some hardware and technical issues have had an impact on quality due to moving from a controlled environment to a real-life environment. The data collection was an uncontrolled, real-life experiment. Some of the most significant challenges were as follows:

- Hardware issues: Internet connection, headset type
- Limited time
- Significant effort
- Limited finances
- Some participants reported a very extreme low point, ignoring instructions for a mild low point. Hence, three audio tracks that included extremely bad experiences were edited to be more vague.
- Uncontrollable environmental noise: street, animals or pets, neighbours, cars, dishwasher, washing machine, going up or down stairs, yelling, screaming, airplane, vacuum, slamming doors, mouse clicks, sniffing, sneezing in the background, coughing, and tongue clicks.

Another challenges was related to the augmentation process that was necessary in this research. When applying SpecAugment, the output was a spectrogram but OpenSMILE only accepts audio files as input. Therefore, after augmenting the audio file, it must be converted back to the WAV format to be inserted into OpenSMILE and proceed with feature extraction.

References

Aarabi, P., Hughes, D., Mohajer, K., & Emami, M. (2001). The automatic measurement of facial beauty. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace* (Vol. 4, p. 2644-2647). IEEE.

Ābele, L., Haustein, S., Møller, M., & Zettler, I. (2020). Links between observed and self-reported driving anger, observed and self-reported aggressive driving, and personality traits. *Accident Analysis & Prevention*, *140*, 105516.

Adler, J. M., Lodi-Smith, J., Philippe, F. L., & Houle, I. (2016). The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review*, *20*(2), 142–175.

Adnan, M., Mukhtar, H., & Naveed, M. (2012). Persuading students for behavior change by determining their personality type. In *2012 15th International Multitopic Conference (INMIC)* (pp. 439–449). IEEE.

Aghaei, M., Parezzan, F., Dimiccoli, M., Radeva, P., & Cristani, M. (2017). Clothing and people-a social signal processing perspective. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 532–537). IEEE.

Akosa, J. (2017). Predictive accuracy: a misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum* (pp. 2–5).

Alam, F., & Riccardi, G. (2013). Comparative study of speaker personality traits recognition in conversational and broadcast news speech..

Alam, F., & Riccardi, G. (2014a). Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 955–959). IEEE.

Alam, F., & Riccardi, G. (2014b). Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (pp. 15–18). ACM.

Allen, B. (2015). *Personality theories: Development, growth, and diversity.* Taylor & Francis.

Allport, G. W. (1937). *Personality: A psychological interpretation.* H. Holt.

Allport, G. W. (1961). *Pattern and growth in personality.*

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological monographs*, *47*(1), i.

Allport, G. W., Vernon, P. E., & Powers, E. (1933). *Studies in expressive movement.* The Macmillan Company.

Al-Samarraie, H., Eldenfria, A., & Dawoud, H. (2017). The impact of personality traits on users' information-seeking behavior. *Information Processing & Management*, *53*(1), 237–247.

Aly, A., Tapus, A., et al. (2012). Robot personality design for an appropriate response to the human partner. In *Feedback Readability for Robots Workshop (in Conjunction with the 21st IEEE International Symposium on Robot and Human Interactive Communication "Ro-Man").*

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, *16*(1), 4–13.

Ambady, N., & Rosenthal, R. (1992). *Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.* American Psychological Association.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, *64*(3), 431.

An, G., Brizan, D. G., & Rosenberg, A. (2013). Detecting laughter and filled pauses using syllable-based features. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 178–181).

An, G., & Levitan, R. (2018). Comparing approaches for mitigating intergroup variability in personality recognition. *arXiv preprint arXiv:1802.01405*.

An, G., Levitan, S. I., Levitan, R., Rosenberg, A., Levine, M., & Hirschberg, J. (2016). Automatically classifying self-rated personality scores from speech. In *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association* (pp. 1412–1416).

Aslan, S., & Güdükbay, U. (2019). Multimodal video-based apparent personality recognition using long short-term memory and convolutional neural networks. *arXiv preprint arXiv:1911.00381*.

Atkinson, R. (1998). *The life story interview.* SAGE Publications. Retrieved from https://books.google.com.sa/books?id=4nwoDAAAQBAJ

Audhkhasi, K., Metallinou, A., Li, M., & Narayanan, S. (2012). Speaker personality classification using systems based on acoustic-lexical cues and an optimal tree-structured bayesian network. In *Thirteenth Annual Conference of the International Speech Communication Association.* Citeseer.

Aufegger, L., Bicknell, C., Soane, E., Ashrafian, H., & Darzi, A. (2019). Understanding health management and safety decisions using signal processing and machine learning. *BMC medical research methodology*, *19*(1), 121.

Aydin, B., Kindiroglu, A. A., Aran, O., & Akarun, L. (2016). Automatic personality prediction from audiovisual data using random forest regression. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (p. 37-42).

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, *124*, 150–159.

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, *18*(1998), 1–8.

Bandura, A., & Walters, R. H. (1977). Social learning theory.

Banko, M., & Brill, E. (2001). Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the First International Conference on Human Language Technology Research.*

Barnard, P., May, J., Duke, D., & Duce, D. (2000, June). Systems, interactions, and macrotheory. *ACM Trans. Comput.-Hum. Interact.*, *7*(2), 222–262.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Batrinca, L., Lepri, B., Mana, N., & Pianesi, F. (2012a). Multimodal recognition of personality traits in human-computer collaborative tasks. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 39–46). New York, NY, USA: ACM.

Batrinca, L., Lepri, B., Mana, N., & Pianesi, F. (2012b). Multimodal recognition of personality traits in human-computer collaborative tasks. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 39–46). ACM.

Batrinca, L., Lepri, B., & Pianesi, F. (2011). Multimodal recognition of personality during short self-presentations. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding* (pp. 27–28). ACM.

Batrinca, L., Mana, N., Lepri, B., Pianesi, F., & Sebe, N. (2011). Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (pp. 255–262). ACM.

Batrinca, L., Mana, N., Lepri, B., Sebe, N., & Pianesi, F. (2016). Multimodal personality recognition in collaborative goal-oriented tasks. *IEEE Transactions on Multimedia*, *18*(4), 659–673.

Bauer, J. J., McAdams, D. P., & Sakaeda, A. R. (2005a). Crystallization of desire and crystallization of discontent in narratives of life-changing decisions. *Journal of personality*, *73*(5), 1181–1214.

Bauer, J. J., McAdams, D. P., & Sakaeda, A. R. (2005b). Interpreting the good life: Growth memories in the lives of mature, happy people. *Journal of personality and social psychology*, *88*(1), 203.

Beer, A., Rogers, K. H., & Letzring, T. D. (2019). Effects of escalated exposure to information on accuracy of personality judgment. *Journal of Research in Personality*, *83*, 103864.

Benet-Martinez, V., & John, O. P. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, *75*(3), 729.

Beyan, C., Zunino, A., Shahid, M., & Murino, V. (2019). Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images. *IEEE Transactions on Affective Computing*.

Biel, J.-I., & Gatica-Perez, D. (2013). The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, *15*(1), 41–55.

Biel, J.-I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013). Hi youtube!: Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (pp. 119–126). ACM.

Blackman, M. C. (2002). Personality judgment and the utility of the unstructured employment interview. *Basic and applied social psychology*, *24*(3), 241–250.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, *34*(2), 164–181.

Blackman, M. C., & Funder, D. C. (2002). Effective interview practices for accurately assessing counterproductive traits. *International Journal of Selection and Assessment*, *10*(1-2), 109–116.

Blalock, E. M. (2003). *A beginner's guide to microarrays*. Springer Science &

Business Media.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj*, *314*(7080), 572.

Boersma, P., & Van Heuven, V. (2001). Speak and unspeak with praat. *Glot International*, *5*(9/10), 341–347.

Bone, D., Chaspari, T., Audhkhasi, K., Gibson, J., Tsiartas, A., Van Segbroeck, M., … Narayanan, S. S. (2013). Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 182–186).

Booth, P. (2014). *An introduction to human-computer interaction (psychology revivals)*. Taylor & Francis.

Booth-Kewley, S., & Vickers, R. R. (1994). Associations between major domains of personality and health behavior. *Journal of personality*, *62*(3), 281–298.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010* (pp. 177–186). Springer.

Bouchard, T. J., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior genetics*, *31*(3), 243–273.

Bousmalis, K., Mehu, M., & Pantic, M. (2009). Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009* (pp. 1–9). IEEE.

Bousmalis, K., Mehu, M., & Pantic, M. (2013). Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, *31*(2), 203–221.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brueckner, R., & Schuller, B. (2013). Hierarchical neural networks and enhanced class posteriors for social signal classification. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 362–367). IEEE.

Brueckner, R., & Schulter, B. (2014). Social signal classification using deep blstm recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4823–4827). IEEE.

Brunet, P. M., Donnan, H., McKeown, G., Douglas-Cowie, E., & Cowie, R. (2009, Sept). Social signal processing: What are the relevant variables? and in what ways do they relate? In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (p. 1-6).

Bucher, M. A., Suzuki, T., & Samuel, D. B. (2019). A meta-analytic review of personality traits and their associations with mental health treatment outcomes. *Clinical psychology review*, *70*, 51–63.

Bühler, J. L., Finkenauer, C., & Grob, A. (2020). A dyadic personality perspective on the michelangelo phenomenon: How personality traits relate to people's ideal selves and their personal growth in romantic relationships. *Journal of Research in Personality*, 103943.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications.* Springer Berlin Heidelberg. Retrieved from `https://books.google.com.sa/books?id=S6jYXmh988UC`

Burger, J. (2014). *Personality.* Cengage Learning.

Cai, R., Guo, A., Ma, J., Huang, R., Yu, R., & Yang, C. (2018). Correlation analyses between personality traits and personal behaviors under specific emotion states using physiological data from wearable devices. In *2018 ieee 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 46–53). IEEE.

Carbonneau, M.-A., Granger, E., Attabi, Y., & Gagnon, G. (2017). Feature learning from spectrograms for assessment of personality traits. *IEEE Transactions on Affective Computing*.

Carducci, B. J. (2009). *The psychology of personality: Viewpoints, research, and applications.* John Wiley & Sons.

Carey, J. (1997). *Relationship between user interface design and human performance.* Intellect, Limited.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*(5), 1054–1072.

Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual review of psychology*, *48*(1), 61–83.

Carroll, J. M. (2003). Chapter 1 - introduction: Toward a multidisciplinary science of human-computer interaction. In J. M. Carroll (Ed.), *HCI Models, Theories, and Frameworks* (p. 1 - 9). San Francisco: Morgan Kaufmann.

Cassell, J. (2000a). Embodied conversational agents. In (pp. 1–27). Cambridge, MA, USA: MIT Press.

Cassell, J. (2000b). *Embodied conversational agents.* MIT Press.

Cattell, R. (1966). *The scientific analysis of personality.* Aldine Pub. Co.

Celiktutan, O., Skordos, E., & Gunes, H. (2019). Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, *10*(4), 484-497.

Celli, F. (2012). Unsupervised personality recognition for social network sites. In *Proc. of Sixth International Conference on Digital Society.* Citeseer.

Celli, F., Bruni, E., & Lepri, B. (2014). Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 1101–1104). New York, NY, USA: ACM.

Celli, F., Lepri, B., Biel, J.-I., Gatica-Perez, D., Riccardi, G., & Pianesi, F. (2014). The workshop on computational personality recognition.

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). Workshop on computational personality recognition (shared task)..

Cerekovic, A., Aran, O., & Gatica-Perez, D. (2016). Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, *8*(3), 382–395.

Chastagnol, C., & Devillers, L. (2012). Personality traits detection using a parallelized modified sffs algorithm. *computing*, *15*, 16.

Chattopadhyay, A., Dahl, D. W., Ritchie, R. J., & Shahin, K. N. (2003). Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology*, *13*(3), 198–204.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, C. W., Wu, C., & Aghajan, H. (2011, March). Real-time social interaction analysis. In *Face and Gesture 2011* (p. 648-648).

Chen, J., Chang, M. C., & Tu, P. (2015, May). A live video analytic system for affect analysis in public space. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 1, p. 1-1).

Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 55–63.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450.

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, *18*(2), 123–149.

Cieslak, D. A., & Chawla, N. V. (2008). Learning decision trees for unbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 241–256). Springer.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of personality and social psychology*, *68*(6), 1152.

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology*, *60*(6), 884.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Costa, P. T., & McCrae, R. R. (1992a). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, *4*(1), 5.

Costa, P. T., & McCrae, R. R. (1992b). *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi): Professional manual.* Psychological Assessment Resources, Incorporated.

Coulter, T. J., Mallett, C. J., Singer, J. A., & Wrzus, C. (2018). A three–domain personality analysis of a mentally tough athlete. *European Journal of Personality*, *32*(1), 6–29.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive aggressive algorithms.

Cristani, M., Murino, V., & Vinciarelli, A. (2010). Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 51–58). IEEE.

Cristani, M., Raghavendra, R., Del Bue, A., & Murino, V. (2013). Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, *100*, 86–97.

Cronbach, L. J. (1955). Processes affecting scores on" understanding of others" and" assumed similarity.". *Psychological bulletin*, *52*(3), 177.

Cunningham, P., & Delany, S. J. (2020). k-nearest neighbour classifiers–. *arXiv preprint arXiv:2004.04523*.

Darbyshire, D., Kirk, C., Wall, H. J., & Kaye, L. K. (2016). Don't judge a (face) book by its cover: Exploring judgement accuracy of others' personality on facebook. *Computers in Human Behavior*, *58*, 380–387.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357–366.

Demetriou, A., Kazi, S., Spanoudis, G., & Makris, N. (2019). Predicting school performance from cognitive ability, self-representation, and personality from primary school to senior high school. *Intelligence*, *76*, 101381.

D'Errico, F., Leone, G., & Poggi, I. (2010). Types of help in the teacher's multimodal behavior. In *International Workshop on Human Behavior Understanding* (pp. 125–139). Springer.

Dicke, A.-L., Lüdtke, O., Trautwein, U., Nagy, G., & Nagy, N. (2012). Judging students' achievement goal orientations: Are teacher ratings accurate? *Learning and Individual Differences*, *22*(6), 844–849.

Dix, A. (2004). *Human-computer interaction.* Pearson/Prentice-Hall.

Dobewall, H., Aavik, T., Konstabel, K., Schwartz, S. H., & Realo, A. (2014). A comparison of self-other agreement in personal values versus the big five personality traits. *Journal of Research in Personality*, *50*, 1–10.

Dong, W., Lepri, B., & Pentland, A. (2012). Automatic prediction of small group performance in information sharing tasks. *arXiv preprint arXiv:1204.3698*.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. (1997). Support vector regression machines. *Advances in neural information processing systems*, *9*, 155–161.

Dudani, S. A. (1976, April). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(4), 325-327.

Eduardo, M.-G., & Ildefonso, M. M. (2020). On the long-run association between personality traits and road crashes: findings from the british cohort study. *Personality and individual differences*, *155*, 109677.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360–392.

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM*

Conference on Information and Knowledge Management (pp. 127–136). ACM.

Escalante, H. J., Ponce-López, V., Wan, J., Riegler, M. A., Chen, B., Clapés, A., … others (2016). Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 67–73). IEEE.

Esposito, A., Esposito, A. M., & Vogel, C. (2015). Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recognition Letters*, *66*, 41–51.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, *20*(1), 18–36.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 835–838). ACM.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1459–1462). ACM.

Eysenck, H. J. (1950). *Dimensions of personality* (Vol. 5). Transaction Publishers.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*.

Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M., & Davalos, S. (2014). A multivariate regression approach to personality impression recognition of vloggers. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (pp. 1–6). ACM.

Farnadi, G., Zoghbi, S., Moens, M.-F., & De Cock, M. (2013). Recognising personality traits using facebook status updates. In *Proceedings of the Workshop on Computational Personality Recognition (WCPR13).* AAAI.

Feldman, R. (2014). *Applications of nonverbal behavioral theories and research.* Taylor & Francis.

Feng, E. L. D., Neo, Z.-W., De Silva, A. W., Sim, K., Tan, H.-R., Nguyen, T.-T., … Nguyen, H. D. (2020). Social behaviour understanding using deep neural networks: Development of social intelligence systems. In *International Conference on Human-Computer Interaction* (pp. 600–613). Springer.

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018, January). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Int. Res.*, *61*(1), 863–905.

Fernández, A., López, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, *42*, 97–110.

Finnerty, A. N., Lepri, B., & Pianesi, F. (2016). Acquisition of personality. In M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, & A. Košir (Eds.), *Emotions and Personality in Personalized Services: Models, Evaluation and Applications* (pp. 81–99). Springer International Publishing.

Flutura, S., Wagner, J., Lingenfelser, F., Seiderer, A., & André, E. (2016). Mobilessi-a multi-modal framework for social signal interpretation on mobile devices. In *2016 12th International Conference on Intelligent Environments (IE)* (pp. 210–213). IEEE.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*(2), 256 - 285.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2-3), 131–163.

Funder, D. C. (1993). Judgments as data for personality and developmental psychology: Error versus accuracy.

Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychological review*, *102*(4), 652.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* Elsevier.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of personality and social psychology*, *55*(1), 149.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of personality and social psychology*, *64*(3), 479.

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of personality*, *61*(4), 457–476.

Galitz, W. (2002). *The essential guide to user interface design: An introduction to gui design principles and techniques.* Wiley.

Galitz, W. (2007). *The essential guide to user interface design: An introduction to gui design principles and techniques.* John Wiley & Sons.

Gan, D.-S. H. Y., & Gromiha, P. G. M. M. (2010). Advanced intelligent computing theories and applications.

Ganguli, R., Mehta, A., & Sen, S. (2020). A survey on machine learning methodologies in social network analysis. In *8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 484–489). IEEE.

Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., & Knoll, A. (2012, Oct). Social behavior recognition using body posture and head pose for human-robot interaction. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (p. 2128-2133).

Gatica-Perez, D. (2014). Signal processing in the workplace [social sciences]. *IEEE Signal Processing Magazine*, *32*(1), 121–125.

Gatica-Perez, D., Vinciarelli, A., & Odobez, J.-M. (2014). Nonverbal behavior analysis. In *Multimodal Interactive Systems Management* (pp. 165–187). EPFL Press.

Gavrilescu, M. (2015a). Study on determining the big-five personality traits of an individual based on facial expressions. In *E-Health and Bioengineering Conference (EHB), 2015* (pp. 1–6). IEEE.

Gavrilescu, M. (2015b, Nov). Study on determining the big-five personality traits of an individual based on facial expressions. In *2015 E-Health and Bioengineering Conference (EHB)* (p. 1-6).

Gievska, S., & Koroveshovski, K. (2014). The impact of affective verbal content on predicting personality impressions in youtube videos. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (p. 19–22). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2659522.2659529` doi: 10.1145/2659522.2659529

Gilal, A. R., Jaafar, J., Basri, S., Omar, M., & Tunio, M. Z. (2015, May). Making programmer suitable for team-leader: Software team composition based on personality types. In *2015 International Symposium on Mathematical Sciences and Computing Research (iSMSC)* (p. 78-82).

Gilpin, L. H., Olson, D. M., & Alrashed, T. (2018). Perception of speaker personality traits using speech signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–6).

Girden, E. R. (1992). *Anova: Repeated measures* (No. 84). Sage.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 1, pp. 517–520). IEEE Computer Society.

Goldberg, L. R. (1980). A catalogue of 1947 nouns that can be used to describe personality and a taxonomy of 1342 nouns that are typically so used. *Unpublished report, Oregon Research Institute.*

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, *2*(1), 141–165.

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, *4*(1), 26.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, *48*(1), 26.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, *40*(1), 84–96.

Goldberg, L. R., & Saucier, G. (1998). What is beyond the big five? *Journal of personality*, *66*(4), 495–524.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge.

Gosztolya, G., Busa-Fekete, R., & Tóth, L. (2013). Detecting autism, emotions and social signals using adaboost. Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association.

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83–90.

Griffin, H. J., Aung, M. S., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., & Bianchi-Berthouze, N. (2013). Laughter type recognition from whole body motion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 349–355). IEEE.

Griffin, H. J., Aung, M. S. H., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., & Bianchi-Berthouze, N. (2015, April). Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives. *IEEE Transactions on Affective Computing*, *6*(2), 165-178.

Gupta, R., Audhkhasi, K., Lee, S., & Narayanan, S. S. (2013). Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 173–177).

Gürpınar, F., Kaya, H., & Salah, A. A. (2016). Combining deep facial and ambient features for first impression estimation. In *European Conference on Computer Vision* (pp. 372–385). Springer.

Gürpinar, F., Kaya, H., & Salah, A. A. (2016). Multimodal fusion of audio, scene, and face features for first impression estimation. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 43–48). IEEE.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389–422.

Halama, P., & Gurnáková, J. (2014). Need for structure and big five personality traits as predictors of decision making styles in health professionals. *Studia Psychologica*, *56*(3), 171.

Hall, J. A., & Goh, J. X. (2017). Studying stereotype accuracy from an integrative social-personality perspective. *Social and Personality Psychology Compass*, *11*(11), e12357.

Hall, J. A., Goh, J. X., Mast, M. S., & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, *84*(4), 433–445.

Hall, J. A., Gunnery, S. D., Letzring, T. D., Carney, D. R., & Colvin, C. R. (2017). Accuracy of judging affect and accuracy of judging personality: How and when are they related? *Journal of Personality*, *85*(5), 583–592.

Han, S., Huang, H., & Tang, Y. (2020). Knowledge of words: An interpretable approach for personality recognition from social media. *Knowledge-Based Systems*, 105550.

Hand, D. J., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern recognition letters*, *24*(9-10), 1555–1562.

Harrigan, J., Rosenthal, R., Scherer, K., & Scherer, K. (2008). *New handbook of methods in nonverbal behavior research.* OUP Oxford.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Hatch, A. O., Kajarekar, S. S., & Stolcke, A. (2006). Within-class covariance normalization for svm-based speaker recognition. In *Proceedings of INTERSPEECH, 7th Annual Conference of the International Speech Communication Association.*

Haylock, M., & Kampkötter, P. (2019). The role of preferences, attitudes, and personality traits in labor market matching. *Economics Letters*, *185*, 108718.

Helander, M., Landauer, T., & Prabhu, P. (1997). *Handbook of human-computer* *interaction*. Elsevier Science.

Hertz, J. A. (2018). *Introduction to the theory of neural computation*. CRC Press.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American*, *267*(3), 144–151.

Hirschmüller, S., Egloff, B., Schmukle, S. C., Nestler, S., & Back, M. D. (2015). Accurate judgments of neuroticism at zero acquaintance: A question of relevance. *Journal of personality*, *83*(2), 221–228.

Hofstee, W. K. (1994). Who should own the definition of personality? *European Journal of Personality*, *8*(3), 149–162.

Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience*, *12*, 105.

Idris, I. (2016). *Python data analysis cookbook*. Packt Publishing. Retrieved from `https://books.google.com.sa/books?id=nZWqDQAAQBAJ`

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies*, *53*(2), 251–267.

Israel, A., Lüdtke, O., & Wagner, J. (2019). The longitudinal association between personality and achievement in adolescence: Differential effects across all big five traits and four achievement indicators. *Learning and Individual Differences*, *72*, 80–91.

Ivanov, A. V., Riccardi, G., Sporka, A. J., & Franc, J. (2011). Recognition of personality traits from human spoken conversations. In *Proceedings of INTERSPEECH, 12th Annual Conference of the International Speech Communication Association* (pp. 1549–1552).

Jang, M., Lee, D. H., Kim, J., & Cho, Y. (2013, Aug). Identifying principal social signals in private student-teacher interactions for robot-enhanced education. In *2013 IEEE RO-MAN* (p. 621-626).

Janicki, A. (2013). Non-linguistic vocalisation recognition based on hybrid gmm-svm approach. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 153–157).

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429–449.

Jerram, K. L., & Coleman, P. G. (1999). The big five personality traits and reporting of health problems and health behaviour in old age. *British Journal of Health Psychology*, *4*(2), 181–192.

Jiang, B., Valstar, M. F., & Pantic, M. (2012, Sept). Facial action detection using block-based pyramid appearance descriptors. In *2012 International Conference* *on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (p. 429-434).

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, *2*(3), 171–203.

John, O. P., Robins, R. W., & Pervin, L. A. (2008). *Handbook of personality, third edition: Theory and research.* Guilford Publications.

John, O. P., Robins, R. W., & Pervin, L. A. (2010). *Handbook of personality: Theory and research.* Guilford Press.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102–138.

Joshi, H., Verma, A., & Mishra, A. (2020). Classification of social signals using deep lstm-based recurrent neural networks. In *2020 International Conference on Signal Processing and Communications (SPCOM)* (pp. 1–5). IEEE.

Joshi, J., Gunes, H., & Goecke, R. (2014). Automatic prediction of perceived traits using visual cues under varied situational context. In *22nd International Conference on Pattern Recognition* (pp. 2855–2860). IEEE.

Jothilakshmi, S., & Brindha, R. (2016). Speaker trait prediction for automatic personality perception using frequency domain linear prediction features. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 2129–2132). IEEE.

Joyce, J. (2003). Bayes' theorem.

Kachur, A., Osin, E., Davydov, D., Shutilov, K., & Novokshonov, A. (2020). Assessing the big five personality traits using real-life static facial images. *Scientific reports*, *10*(1), 1–11.

Kalayci, S., Ekenel, H. K., & Gunes, H. (2014). Automatic analysis of facial attractiveness from video. In *2014 IEEE International Conference onImage Processing (ICIP)* (pp. 4191–4195). IEEE.

Kamisaka, D., & Ishikawa, Y. (2020). Multimodal self-assessed personality prediction in the wild. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (pp. 57–61).

Kasano, E., Muramatsu, S., Matsufuji, A., Sato-Shimokawara, E., & Yamaguchi, T. (2019). Estimation of speaker's confidence in conversation using speech information and head motion. In *2019 16th International Conference on Ubiquitous Robots (UR)* (pp. 294–298). IEEE.

Kaushal, V., & Patwardhan, M. (2018). Emerging trends in personality identification using online social networks—a literature survey. *ACM Transactions on*

Knowledge Discovery from Data (TKDD), *12*(2), 1–30.

Kern, A., Kramm, C., Witt, C. M., & Barth, J. (2020). The influence of personality traits on the placebo/nocebo response: A systematic review. *Journal of Psychosomatic Research*, *128*, 109866.

Kim, S., Filippone, M., Valente, F., & Vinciarelli, A. (2012). Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 793–796).

Kim, S., Valente, F., Filippone, M., & Vinciarelli, A. (2014, April). Predicting continuous conflict perception with bayesian gaussian processes. *IEEE Transactions on Affective Computing*, *5*(2), 187-200.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137–163.

Kirchhoff, K., Liu, Y., & Bilmes, J. A. (2013). Classification of developmental disorders from speech signals using submodular feature selection. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 187–190).

Klados, M. A., Konstantinidi, P., Dacosta-Aguayo, R., Kostaridou, V.-D., Vinciarelli, A., & Zervakis, M. (2020). Automatic recognition of personality profiles using eeg functional connectivity during emotional processing. *Brain sciences*, *10*(5), 278.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Springer.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of personality*, *64*(2), 311–337.

Komarraju, M., & Karau, S. J. (2005). The relationship between the big five personality traits and academic motivation. *Personality and individual differences*, *39*(3), 557–567.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *European Conference on Machine Learning* (pp. 171–182). Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163.

Kotsiantis, S., & Pintelas, P. (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, *1*(1), 46–55.

Krikke, T. F., & Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 163–167).

Krzyzaniak, S. L., Colman, D. E., Letzring, T. D., McDonald, J. S., Biesanz, J. C., & Back, M. (2019). The effect of information quantity on distinctive accuracy and normativity of personality trait judgments. *European Journal of Personality*, *33*(2), 197–213.

Krzyzaniak, S. L., & Letzring, T. D. (2019). Characteristics of traits that are related to accuracy of personality judgments. In *The Oxford Handbook of Accurate Personality Judgment.*

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML* (Vol. 97, pp. 179–186). Nashville, USA.

Kwiatkowska, M. M., & Rogoza, R. (2019). A modest proposal to link shyness and modesty: Investigating the relation within the framework of big five personality traits. *Personality and Individual Differences*, *149*, 8–13.

Lango, M. (2019). Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study. *Foundations of Computing and Decision Sciences*, *44*(2), 151–178.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction.* John Wiley & Sons.

Lee, K. M., Peng, W., Jin, S.-A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, *56*(4), 754–772.

Lehmann-Willenbrock, N., Hung, H., & Keyton, J. (2017). New frontiers in analyzing dynamic group interactions: Bridging social and computer science. *Small group research*, *48*(5), 519–531.

Leone, G., Migliorisi, S., & Sessa, I. (2016, Oct). Detecting social signals of honesty and fear of appearing deceitful: A methodological proposal. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (p. 000289-000294).

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. *Computers as cognitive tools*, 75–105.

Lepri, B., Kalimeri, K., & Pianesi, F. (2010). Honest signals and their contribution to the automatic analysis of personality traits – a comparative study. In A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human Behavior Understanding: First International Workshop, HBU 2010, Istanbul, Turkey,* *August 22, 2010. Proceedings* (pp. 140–150). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lepri, B., Subramanian, R., Kalimeri, K., Staiano, J., Pianesi, F., & Sebe, N. (2012). Connecting meeting behavior with extraversion—a systematic study. *IEEE Transactions on Affective Computing*, *3*(4), 443–455.

Letzring, T. D. (2010). The effects of judge-target gender and ethnicity similarity on the accuracy of personality judgments. *Social Psychology*.

Letzring, T. D., Colman, D. E., Krzyzaniak, S. L., & Roberts, B. W. (2020). Realistic accuracy model. *The Wiley Encyclopedia of Personality and Individual Differences: Models and Theories*, 341–349.

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of personality and social psychology*, *91*(1), 111.

Leutner, F., Ahmetoglu, G., Akhtar, R., & Chamorro-Premuzic, T. (2014). The relationship between the entrepreneurial personality and the big five personality traits. *Personality and individual differences*, *63*, 58–63.

Lievens, F., De Fruyt, F., & Van Dam, K. (2001). Assessors' use of personality traits in descriptions of assessment centre candidates: A five-factor model perspective. *Journal of Occupational and Organizational Psychology*, *74*(5), 623–636.

Lippmann, R. P. (1987, Apr). An introduction to computing with neural nets. *IEEE ASSP Magazine*, *4*(2), 4-22.

Liu, G., Mercer, T. R., Shearwood, A.-M. J., Siira, S. J., Hibbs, M. E., Mattick, J. S., ... Filipovska, A. (2013). Mapping of mitochondrial rna-protein interactions by digital rnase footprinting. *Cell reports*, *5*(3), 839–848.

Liu, W., & Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 345–356). Springer.

Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2003). Intelligence,"big five" personality traits, and work drive as predictors of course grade. *Personality and Individual Differences*, *35*(6), 1231–1239.

MacKenzie, I. (2012). *Human-computer interaction: An empirical research perspective*. Elsevier Science.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, *30*, 457–500.

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017, March). Deep learning-based document modeling for personality detection from text. *IEEE*

Intelligent Systems, *32*(2), 74–79.

Martin, A., & Eastman, D. (1996). *The user interface design book for the applications programmer.* J. Wiley.

Martinez, B., Valstar, M. F., Jiang, B., & Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*.

Matz, S., Chan, Y. W. F., & Kosinski, M. (2016). Models of personality. In *Emotions and Personality in Personalized Services* (pp. 35–54). Springer.

Mayer, J. D. (2007). Asserting the definition of personality. *The online newsletter for personality science*(1).

Mayer, R. (1975). Technical report no, 37 towards a cognitive model of programmer behavior ben shneiderman richard mayer august, 1975.

McAdams, D. P. (2001). The psychology of life stories. *Review of general psychology*, *5*(2), 100–122.

McAdams, D. P. (2012). Exploring psychological themes through life-narrative accounts. *Varieties of narrative analysis*, 15–32.

McAdams, D. P., Anyidoho, N. A., Brown, C., Huang, Y. T., Kaplan, B., & Machado, M. A. (2004). Traits and stories: Links between dispositional and narrative features of personality. *Journal of personality*, *72*(4), 761–784.

McAdams, D. P., & Pals, J. L. (2006). A new big five: fundamental principles for an integrative science of personality. *American psychologist*, *61*(3), 204.

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? In *Proceedings of the 1st International Workshop on Utility-based Data Mining* (pp. 69–77). ACM.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

McCrae, R. R., & Costa, P. T. (1995). Trait explanations in personality psychology. *European Journal of Personality*, *9*(4), 231–252.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175–215.

McKeown, G., Curran, W., McLoughlin, C., Griffin, H. J., & Bianchi-Berthouze, N. (2013, April). Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (p. 1-5).

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, *90*(5), 862.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods*, *33*(4), 517–523.

Melanie, M. (1996). An introduction to genetic algorithms.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*(1), 213.

Meyer, J., Fleckenstein, J., Retelsdorf, J., & Köller, O. (2019). The relationship of personality traits and different measures of domain-specific achievement in upper secondary education. *Learning and Individual Differences*, *69*, 45–59.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop* (pp. 41–48). Ieee.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mischel, W. (2013). *Personality and assessment.* Psychology Press.

Miwa, H., Umetsu, T., Takanishi, A., & Takanobu, H. (2001). Robot personality based on the equations of emotion defined in the 3d mental space. In *Proceedings 2001 ICRA IEEE International Conference on Robotics and Automation, 2001.* (Vol. 3, pp. 2602–2607). IEEE.

Mohammadi, G., Origlia, A., Filippone, M., & Vinciarelli, A. (2012). From speech to personality: Mapping voice quality and intonation into personality differences. In *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 789–792). ACM.

Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, *3*(3), 273–284.

Mohammadi, G., Vinciarelli, A., & Mortillaro, M. (2010). The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd International Workshop on Social Signal Processing* (pp. 17–20). ACM.

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, *74*, 16–22.

Moreno, A. (2012, Sept). Automatic detection of social signals in digital playgrounds. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (p. 981-982).

Moreno-Armendáriz, M. A., Martínez, C. A. D., Calvo, H., & Moreno-Sotelo, M. (2020). Estimation of personality traits from portrait pictures using the five-factor model. *IEEE Access*, *8*, 201649–201665.

Moritz, D., & Roberts, J. E. (2018). Self-other agreement and metaperception accuracy across the big five: Examining the roles of depression and self-esteem. *Journal of personality*, *86*(2), 296–307.

Myers, D. (2004). *Psychology, seventh edition, in modules (spiral)*. Worth Publishers.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 1, pp. 1–6). IEEE.

Najjab, A., Palka, J. M., & Brown, E. S. (2020). Personality traits and risk of lifetime asthma diagnosis. *Journal of Psychosomatic Research*, *131*, 109961.

Nasir, M., Baucom, B. R., Georgiou, P., & Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PloS one*, *12*(9), e0185123.

Nass, C., & Brave, S. (2007). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81–103.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). ACM.

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, *35*(12), 1661–1671.

Navarathna, R., Carr, P., Lucey, P., & Matthews, I. (2017). Estimating audience engagement to predict movie ratings. *IEEE Transactions on Affective Computing*, *10*(1), 48–59.

Nielsen, J. (1989). *Coordinating user interfaces for consistency*. Morgan Kaufmann.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574.

Nowson, S., & Gill, A. J. (2014). Look! who's talking?: Projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (pp. 23–26). ACM.

O'Connor, M. C., & Paunonen, S. V. (2007). Big five personality predictors of post-secondary academic performance. *Personality and Individual differences*, *43*(5), 971–990.

Oh, J., Cho, E., & Slaney, M. (2013). Characteristic contours of syllabic-level units in laughter. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 158–162).

Okwechime, D., Ong, E. J., Gilbert, A., & Bowden, R. (2011, March). Visualisation and prediction of conversation interest through mined social signals. In *Face and Gesture 2011* (p. 951-956).

Olguín, D., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(1), 43–55.

Olguín, D. O. (2007). *Sociometric badges: wearable technology for measuring human behavior* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Olguin, D. O., Paradiso, J. A., & Pentland, A. (2006). Wearable communicator badge: Designing a new platform for revealing organizational dynamics. In *Proceedings of the 10th International Symposium on Wearable Computers (Student Colloquium)* (pp. 4–6).

Palomba, A. (2020). Consumer personality and lifestyles at the box office and beyond: How demographics, lifestyles and personalities predict movie consumption. *Journal of Retailing and Consumer Services*, *55*, 102083.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). IEEE.

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., ... Vinciarelli, A. (2011). Social signal processing: the research agenda. In *Visual Analysis of Humans* (pp. 511–538). Springer.

Pantic, M., & Vinciarelli, A. (2014). Social signal processing. *The Oxford Handbook of Affective Computing*, 84.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Paunonen, S. V., & Hong, R. Y. (2013). The many faces of assumed similarity in perceptions of personality. *Journal of Research in Personality*, *47*(6), 800–815.

Pease, C. R., & Lewis, G. J. (2015). Personality links to anger: Evidence for trait interaction and differentiation across expression style. *Personality and Individual Differences*, *74*, 159–164.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pejovic, V., & Musolesi, M. (2015). Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Computing Surveys (CSUR)*, *47*(3), 1–29.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Pentland, A. (2007a). Automatic mapping and modeling of human networks. *Physica A: Statistical Mechanics and its Applications*, *378*(1), 59–67.

Pentland, A. (2007b, July). Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine*, *24*(4), 108-111.

Pentland, A., & Heibeck, T. (2010). *Honest signals: How they shape our world.* MIT Press.

Pervin, L. A., Cervone, D., & John, O. P. (2005). *Personality: theory and research.* Wiley.

Pianesi, F. (2013, Jan). Searching for personality [social sciences]. *IEEE Signal Processing Magazine*, *30*(1), 146-158.

Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., & Zancanaro, M. (2008). Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (pp. 53–60).

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Poggi, I., & D'Errico, F. (2012). Social signals: a framework in terms of goals and beliefs. *Cognitive Processing*, *13*(2), 427–445.

Poggi, I., & Francesca, D. (2010). Cognitive modelling of human social signals. In *Proceedings of the 2nd International Workshop on Social Signal Processing* (pp. 21–26). ACM.

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language*, *29*(1), 145–171.

Polychroniou, A., Salamin, H., & Vinciarelli, A. (2014). The sspnet-mobile corpus: Social signal processing over mobile phones. In *LREC* (pp. 1492–1498).

Polzehl, T., Moller, S., & Metze, F. (2010). Automatically assessing personality from speech. In *2010 IEEE Fourth International Conference on Semantic Computing (ICSC)* (pp. 134–140). IEEE.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... Escalera, S. (2016a). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision* (pp. 400–418). Springer.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... Escalera, S. (2016b). Chalearn lap 2016: First round challenge on first impressions - dataset and results. In G. Hua & H. Jégou (Eds.), *Computer Vision – ECCV 2016 Workshops* (pp. 400–418). Springer International Publishing.

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. (2013). Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence* (pp. 484–496). Springer.

Preece, J., Sharp, H., & Rogers, Y. (2015). *Interaction design: Beyond human-computer interaction.* Wiley.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, *15*(11), 1119–1125.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning.* Elsevier.

Rabiner, L., & Juang, B.-H. (1986). An introduction to hidden markov models. *ieee assp magazine*, *3*(1), 4–16.

Raggatt, P. (2006). Putting the five-factor model into context: Evidence linking big five traits to narrative identity. *Journal of Personality*, *74*(5), 1321–1348.

Ragni, A., Knill, K., Rath, S. P., & Gales, M. (2014). Data augmentation for low resource languages.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, *41*(1), 203–212.

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, *27*(1), 89–115.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4-21.

Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self-and peer-reported personality: Agreement, trait ratability, and the "self-based heuristic". *Journal of Research in Personality*, *34*(2), 208–224.

Reisz, Z., Boudreaux, M. J., & Ozer, D. J. (2013). Personality traits and the prediction of personal goals. *Personality and Individual Differences*, *55*(6), 699–704.

Reynolds, D. (2015). Gaussian mixture models. *Encyclopedia of biometrics*, 827–832.

Ridgell, S. D., & Lounsbury, J. W. (2004). Predicting academic success: General intelligence," big five" personality traits, and work drive. *College Student Journal*, *38*(4), 607–619.

Riggio, R., & Feldman, R. (2005). *Applications of nonverbal communication.* Taylor & Francis.

Rissola, E. A., Bahrainian, S. A., & Crestani, F. (2019). Personality recognition in conversations using capsule neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 180–187).

Rogers, Y. (2004). New theoretical approaches for human-computer interaction. *Annual Review of Information Science and Technology*, *38*(1), 87–143.

Rogers, Y. (2012). *Hci theory: Classical, modern, and contemporary.* Morgan & Claypool.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Roudposhti, K. K., Nunes, U., & Dias, J. (2015). Probabilistic social behavior analysis by exploring body motion-based patterns. *IEEE transactions on pattern analysis and machine intelligence*, *38*(8), 1679–1691.

Rubenstein, A. L., Zhang, Y., Ma, K., Morrison, H. M., & Jorgensen, D. F. (2019). Trait expression through perceived job characteristics: A meta-analytic path model linking personality and job attitudes. *Journal of Vocational Behavior*, *112*, 141–157.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660–674.

Sain, H., & Purnami, S. W. (2015). Combine sampling support vector machine for imbalanced data classification. *Procedia Computer Science*, *72*, 59–66.

Salah, A. A., Pantic, M., & Vinciarelli, A. (2011). Recent developments in social signal processing. In *2011 IEEE International Conference on Systems, Man,* *and Cybernetics (SMC)* (pp. 380–385). IEEE.

Salamin, H., Polychroniou, A., & Vinciarelli, A. (2013). Automatic recognition of personality and conflict handling style in mobile phone conversations. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (pp. 1–4). IEEE.

Sanchez-Cortes, D., Aran, O., & Gatica-Perez, D. (2011). An audio visual corpus for emergent leader analysis. In *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road Mapping the Future, ICMI-MLMI.*

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in aging neuroscience*, *9*, 329.

Sarkar, C., Bhatia, S., Agarwal, A., & Li, J. (2014). Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (pp. 11–14). ACM.

Sarle, W. S. (1994). Neural networks and statistical models.

Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine learning*, *39*(2-3), 135–168.

Schmid Mast, M., Bangerter, A., Bulliard, C., & Aerni, G. (2011). How accurate are recruiters' first impressions of applicants in employment interviews? *International Journal of Selection and Assessment*, *19*(2), 198–208.

Schötz, S. (2002). Linguistic & paralinguistic phonetic variation in speaker recognition & text-to-speech synthesis. Citeseer.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., ... others (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association* (pp. 2001–2005).

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., ... others (2012). The interspeech 2012 speaker trait challenge. In *Proceedings of INTERSPEECH, 13th Annual Conference of the International Speech Communication Association.*

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., ... others (2015). A survey on perceived speaker traits: personality, likability, pathology, and the first challenge. *Computer Speech & Language*, *29*(1), 100–131.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... others (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association.*

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., ... Lepri, B. (2017). What your facebook profile picture reveals about your personality. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 460–468).

Shneiderman, B. (1979, Dec). Human factors experiments in designing interactive systems. *Computer*, *12*(12), 9-19.

Shneiderman, B. (1981). Putting the human factor into systems development. In *Proceedings of the Eighteenth Annual Computer Personnel Research Conference* (pp. 1–13). New York, NY, USA: ACM.

Shneiderman, B. (1983). Human factors of interactive software. In A. Blaser & M. Zoeppritz (Eds.), *Enduser Systems and their Human Factors: Proceedings of the Scientific Symposium conducted on the occasion of the 15th Anniversary of the Science Center Heidelberg of IBM Germany Heidelberg, March 18, 1983* (pp. 9–29). Berlin, Heidelberg: Springer Berlin Heidelberg.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer-interaction* (No. v. 85). Addison Wesley Longman.

Shneiderman, B., & Plaisant, C. (2003). *Designing the user interface: Fourth edition preview.* Addison Wesley Longman.

Shneiderman, B., & Plaisant, C. (2010). *Designing the user interface: Strategies for effective human-computer interaction.* Addison-Wesley.

Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2017). *Designing the user interface: Strategies for effective human-computer interaction.* Pearson Education.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

Siegert, I., Haase, M., Prylipko, D., & Wendemuth, A. (2014). Discourse particles and user characteristics in naturalistic human-computer interaction. In M. Kurosu (Ed.), *Proceedings of Human-Computer Interaction. Advanced Interaction Modalities and Techniques: 16th International Conference, HCI International 2014* (pp. 492–501). Springer International Publishing.

Sigurdsson, J. F. (1991). Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences*, *12*(6), 617 - 624.

Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*

(Vol. 3).

Staiano, J., Lepri, B., Subramanian, R., Sebe, N., & Pianesi, F. (2011). Automatic modeling of personality states in small group interactions. In *Proceedings of the 19th ACM International Conference on Multimedia* (pp. 989–992). ACM.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, *9*(1), 319.

Stillwell, D. J., & Kosinski, M. (2004). mypersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist*, *59*(2), 93–104.

Stoll, G., Einarsdóttir, S., Song, Q. C., Ondish, P., Sun, J.-T., & Rounds, J. (2020). The roles of personality traits and vocational interests in explaining what people want out of life. *Journal of Research in Personality*, 103939.

Suen, H.-Y., Hung, K.-E., & Lin, C.-L. (2019). Tensorflow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access*, *7*, 61018–61023.

Sutić, D., Brešković, I., Huić, R., & Jukić, I. (2010). Automatic evaluation of facial attractiveness. In *2010 Proceedings of the 33rd International Convention MIPRO* (pp. 1339–1342). IEEE.

Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, *15*(3), 142–147.

Swann Jr, W. B., & Seyle, C. (2005). Personality psychology's comeback and its emerging symbiosis with social psychology. *Personality and Social Psychology Bulletin*, *31*(2), 155–165.

Syrdal, D. S., Koay, K. L., Walters, M. L., & Dautenhahn, K. (2007). A personalized robot companion?-the role of individual differences on spatial preferences in hri scenarios. In *The 16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007.* (pp. 1143–1148). IEEE.

Taber, L., & Whittaker, S. (2018). Personality depends on the medium: differences in self-perception on snapchat, facebook and offline. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Taggart, T. C., Bannon, S. M., & Hammett, J. F. (2019). Personality traits moderate the association between conflict resolution and subsequent relationship satisfaction in dating couples. *Personality and Individual Differences*, *139*, 281–289.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., & Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE journal of translational engineering in health and medicine*, *5*, 1–11.

Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y. L., et al. (2017). Personality prediction system from facebook users. *Procedia computer science*, *116*, 604–611.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478458` doi: 10.1080/01621459.1987.10478458

Tantithamthavorn, C., Hassan, A. E., & Matsumoto, K. (2018). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*.

Tapus, A., & Mataric, M. J. (2008). Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior* (pp. 133–140).

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International journal of medical education*, *2*, 53.

Ter Laak, J. J., De Goede, M. P., & Brugman, G. M. (2001). Teachers'judgments of pupils: Agreement and accuracy. *Social Behavior and Personality: an international journal*, *29*(3), 257–270.

Thomsen, D. K., Olesen, M. H., Schnieber, A., & Tønnesvang, J. (2014). The emotional content of life stories: Positivity bias and relation to personality. *Cognition & emotion*, *28*(2), 260–277.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225–1231.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* (Tech. Rep.). DTIC Document.

Uddin, M. F., & Lee, J. (2016, Aug). Predicting good fit students by correlating relevant personality traits with academic/career data. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (p. 968-975).

Valente, F., Kim, S., & Motlicek, P. (2012a). Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Proceedings of INTERSPEECH, 13th Annual Conference of the International Speech Communication Association.*

Valente, F., Kim, S., & Motlicek, P. (2012b). Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus..

Varni, G., Camurri, A., Coletta, P., & Volpe, G. (2009, Aug). Toward a real-time automated measure of empathy and dominance. In *2009 International Conference on Computational Science and Engineering* (Vol. 4, p. 843-848).

Vazire, S. (2010). Who knows what about a person? the self–other knowledge asymmetry (soka) model. *Journal of personality and social psychology*, *98*(2), 281.

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: the accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of personality and social psychology*, *95*(5), 1202.

Ventura, C., Masip, D., & Lapedriza, A. (2017). Interpreting cnn models for apparent personality trait regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 55–63).

Verhoeven, B., Daelemans, W., & De Smedt, T. (2013). Ensemble methods for personality recognition. In *Proceedings of the Workshop on Computational Personality Recognition(WCPR13)* (pp. 35–38). AAAI.

Verhoeven, B., Soler Company, J., & Daelemans, W. (2014). Evaluating content-independent features for personality recognition. In *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition* (pp. 7–10).

Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–4). IEEE.

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., ... others (2015). Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, *7*(4), 397–413.

Vinciarelli, A., & Mohammadi, G. (2014a). More personality in personality computing. *IEEE Transactions on Affective Computing*, *5*(3), 297–300.

Vinciarelli, A., & Mohammadi, G. (2014b, July). A survey of personality computing. *IEEE Transactions on Affective Computing*, *5*(3), 273-291.

Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008a). Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM International Conference on Multimedia* (pp. 1061–1070). ACM.

Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008b). Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th*

International Conference on Multimodal Interfaces (pp. 61–68). ACM.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), 69–87.

Vinciarelli, A., Salamin, H., & Pantic, M. (2009). Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 42–49). IEEE.

Vinciarelli, A., Salamin, H., Polychroniou, A., Mohammadi, G., & Origlia, A. (2012). From nonverbal cues to perception: personality and social attractiveness. In *Cognitive Behavioural Systems* (pp. 60–72). Springer.

Vinciarelli, A., Valente, F., Yella, S. H., & Sapru, A. (2011). Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus. In *2011 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 374–379). IEEE.

Vogt, D. S., & Randall Colvin, C. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of personality*, *71*(2), 267–295.

Wache, J., Subramanian, R., Abadi, M. K., Vieriu, R.-L., Sebe, N., & Winkler, S. (2015). Implicit user-centric personality recognition based on physiological responses to emotional videos. In *Proceedings of the 2015 ACM on International conference on multimodal interaction* (pp. 239–246).

Wagner, J., Lingenfelser, F., & André, E. (2012). A frame pruning approach for paralinguistic recognition tasks..

Wagner, J., Lingenfelser, F., & André, E. (2013). Using phonetic patterns for detecting social cues in natural conversations. In *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association* (pp. 168–172).

Wei, H., Zhang, F., Yuan, N. J., Cao, C., Fu, H., Xie, X., … Ma, W.-Y. (2017). Beyond the words: Predicting user personality from heterogeneous information. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 305–314).

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., & Scherer, K. (2013a). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, *4*, 292. Retrieved from `https://www.frontiersin.org/article/10.3389/fpsyg.2013.00292` doi: 10.3389/fpsyg.2013.00292

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013b).

On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, *4*, 292.

Westen, D., Gabbard, G. O., & Ortigo, K. M. (1990). Psychoanalytic approaches to personality. *Handbook of personality: Theory and research*, 21–65.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), 826–833.

Whelan, S., & Davies, G. (2006). Profiling consumers of own brands and national brands using human personality. *Journal of Retailing and Consumer Services*, *13*(6), 393–402.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment* (Vol. 8642). Krieger Publishing Company.

Wiggins, J. S. (1996). *The five-factor model of personality: Theoretical perspectives.* Guilford Press.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Wong, S. C., Gatt, A., Stamatescu, V., & McDonnell, M. D. (2016). Understanding data augmentation for classification: when to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–6). IEEE.

Woods, S., Dautenhahn, K., Kaouri, C., Boekhorst, R., & Koay, K. L. (2005). Is this robot like me? links between human and robot personality traits. In *2005 5th IEEE-RAS International Conference on Humanoid Robots* (pp. 375–380). IEEE.

Wright, P., & McCarthy, J. (2008). Empathy and experience in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 637–646). New York, NY, USA: ACM.

Xu, R., Frey, R. M., Fleisch, E., & Ilic, A. (2016). Understanding the impact of personality traits on mobile app adoption–insights from a large-scale field study. *Computers in Human Behavior*, *62*, 244–256.

Yaeger, L. S., Lyon, R. F., & Webb, B. J. (1997). Effective training of a neural network character classifier for word recognition. In *Advances in Neural Information Processing Systems* (pp. 807–816).

Yakasai, A. M., & Jan, M. T. (2015). The impact of big five personality traits on salespeople's performance: Exploring the moderating role of culture. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *4*(5), 11.

Yang, H.-C., & Huang, Z.-R. (2019). Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, *165*, 157–168.

Yang, J., Roy, A., & Zhang, Y. (2013). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, *29*(20), 2588–2595.

Yang, K., Mall, S., & Glaser, N. (2017). *Prediction of personality first impressions with deep bimodal lstm* (Tech. Rep.). Technical report, arXiv, 2017. URL http://cs231n. stanford. edu/reports/2017 ....

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization..

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040.

Yu, J., & Markov, K. (2017). Deep learning based personality recognition from facebook status updates. In *2017 ieee 8th International Conference on Awareness Science and Technology (iCAST)* (pp. 383–387). IEEE.

Yu, Y.-C. (2016). Teaching with a dual-channel classroom feedback system in the digital classroom environment. *IEEE Transactions on Learning Technologies*, *PP*(99), 1-1.

Yuan, C., Wu, J., Li, H., & Wang, L. (2018). Personality recognition based on user generated content. In *2018 15th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1–6). IEEE.

Zhao, J., Zeng, D., Xiao, Y., Che, L., & Wang, M. (2020). User personality prediction based on topic preference and sentiment analysis using lstm model. *Pattern Recognition Letters*, *138*, 397–402.

Zhu, H., Li, L., Zhao, S., & Jiang, H. (2018). Evaluating attributed personality traits from scene perception probability. *Pattern Recognition Letters*, *116*, 121–126.

Zohuri, B., & Moghaddam, M. (2020, 01). Deep learning limitations and flaws.

Appendix A: SPC Experiments Figures and Tables

This appendix includes the tables and figures derived from the Speaker Personality Corpus experiments.

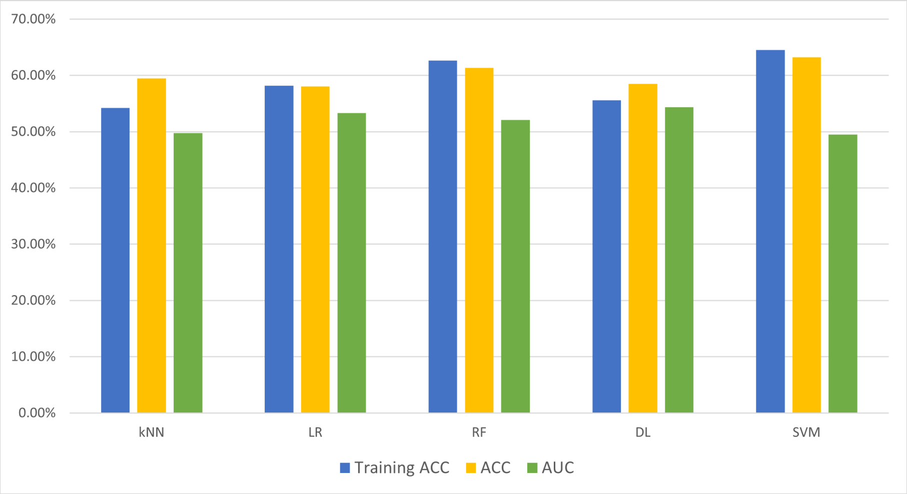


Figure A.1: *Openness trait accuracy measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.1: *Openness trait score based on accuracy fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	54.22%	59.43%	49.73%	49.72%	0.963	0.963	-0.0068
LR	58.16%	58.02%	53.34%	53.33%	0.9864	0.4198	0.0674
RF	62.62%	61.32%	52.12%	52.11%	0.889	0.3867	0.0515
DL	55.59%	58.49%	54.33%	54.32%	0.9806	0.415	0.0865
SVM	64.49%	**63.21%**	49.49%	49.49%	0.7429	0.3679	-0.0228

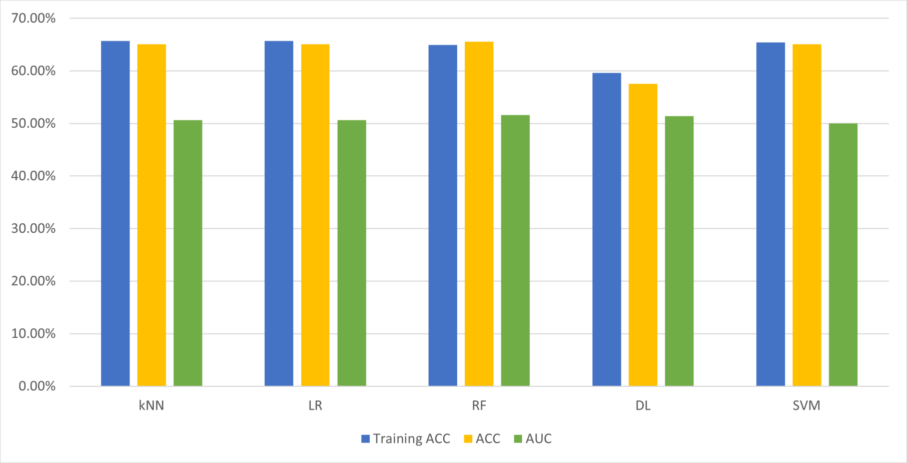


Figure A.2: *Openness trait accuracy measure and parameter tuning is applied to the machine learning algorithms.*

Table A.2: *Openness traits scores based on accuracy fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	65.67%	65.09%	50.63%	50.62%	0.5315	0.349	0.0439
LR	65.67%	65.09%	50.63%	50.62%	0.5315	0.349	0.0439
RF	64.96%	**65.57%**	51.62%	51.61%	0.4741	0.3443	0.0862
DL	59.61%	57.55%	51.41%	51.41%	0.9905	0.4245	0.0295
SVM	65.42%	65.09%	50.00%	50%	0.5315	0.349	None

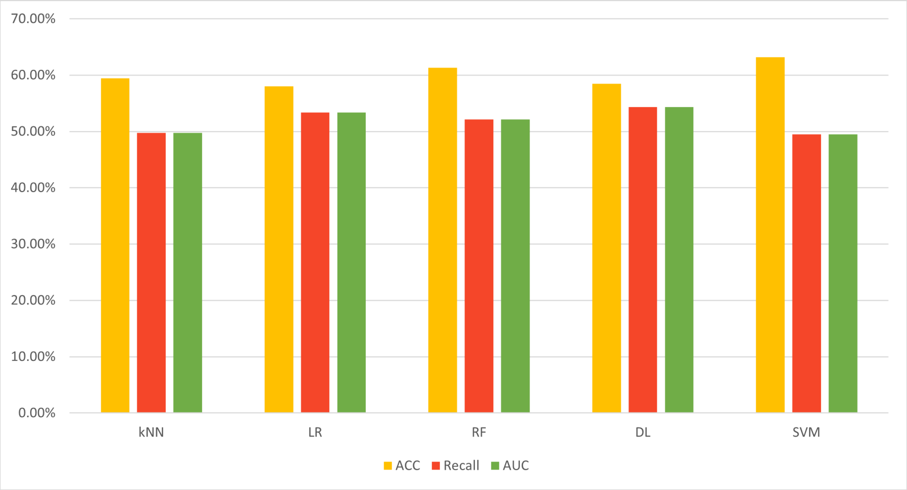


Figure A.3: *Openness trait recall measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.3: *Openness trait scores based on recall fit and no parameter tuning is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	47.78%	59.43%	49.73%	49.72%	0.963	0.4056	-0.0068
LR	50.73%	58.02%	53.34%	53.33%	0.9864	0.4198	0.0674
RF	50.50%	61.32%	52.12%	52.11%	0.889	0.3867	0.0515
DL	52.82%	58.49%	**54.33%**	54.32%	0.9806	0.415	0.0865
SVM	49.51%	63.21%	49.49%	49.49%	0.7429	0.3679	-0.0228

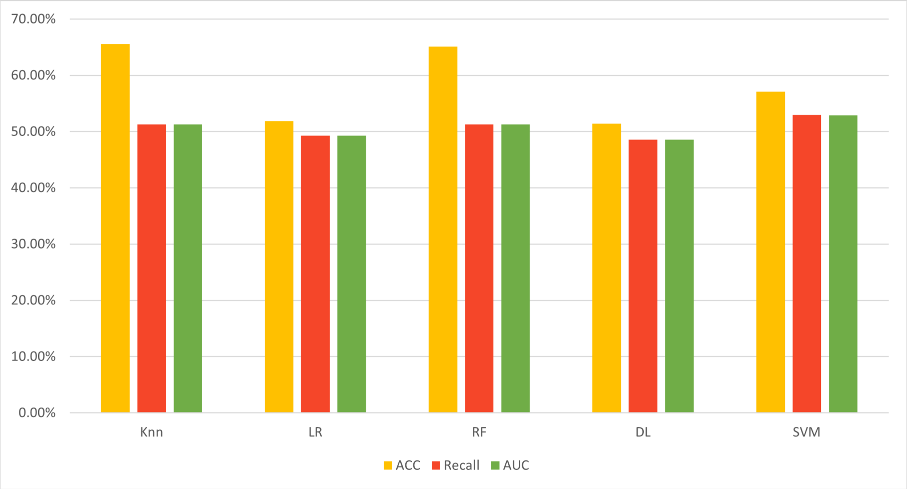


Figure A.4: *Openness trait recall measure and parameter tuning is applied to the machine learning algorithms.*

Table A.4: *Openness trait scores based on recall fit and parameter tuning is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	50.44%	65.57%	51.30%	51.30%	0.4741	0.3443	0.0818
LR	55.58%	51.89%	49.26%	49.25%	0.9999	0.4811	-0.0143
RF	49.93%	65.09%	51.25%	51.25%	0.5315	0.349	0.0627
DL	53.64%	51.42%	48.58%	48.58%	0.9999	0.4858	-0.0275
SVM	50.01%	57.08%	**52.93%**	52.92%	0.9935	0.4292	0.0583

Table A.5: *Openness trait scores based on recall fit and ANOVA as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.61%	50.47%	47.91%	47.90%	0.9996	0.4952	-0.0415
LR	67.91%	54.72%	51.58%	51.57%	0.9851	0.4528	0.0318
RF	51.98%	61.79%	51.18%	51.17%	0.5303	0.382	0.0539
DL	67.21%	54.25%	**52.61%**	52.61%	0.9896	0.4575	0.0513
SVM	66.33%	53.77%	50.58%	50.57%	0.9928	0.4622	0.0117

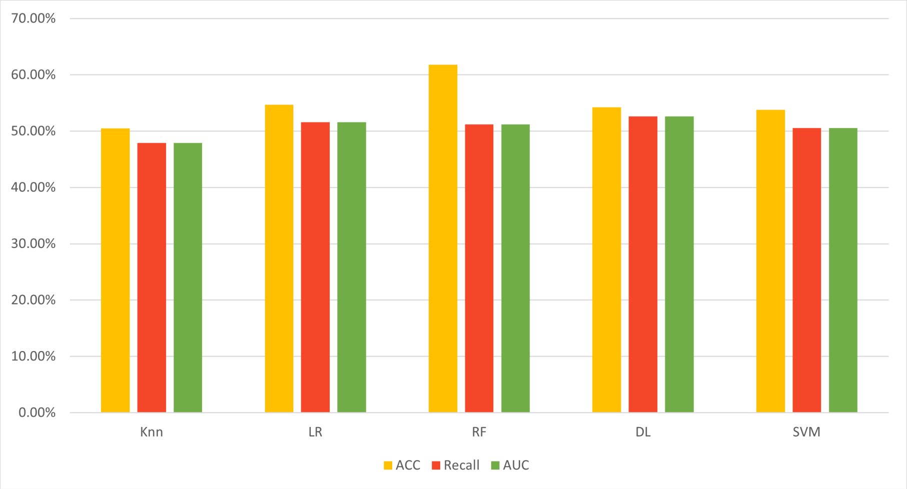


Figure A.5: *Openness trait recall measure and ANOVA as feature reduction is applied to the machine learning algorithms.*

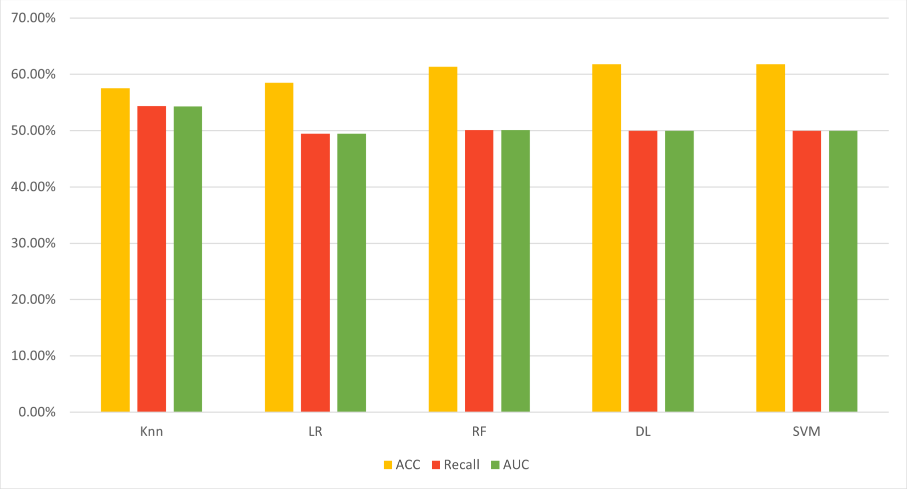


Figure A.6: *Openness trait recall measure and LASSO as feature reduction is applied to the machine learning algorithms.*

Table A.6: *Openness trait scores based on recall fit and LASSO as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	47.43%	57.55%	**54.34%**	54.33%	0.9096	0.4245	0.0882
LR	47.43%	58.49%	49.45%	49.44%	0.8553	0.415	-0.0166
RF	51.42%	61.32%	50.09%	50.08%	0.5859	0.3867	0.0057
DL	50.00%	61.79%	50.00%	50.00%	0.5303	0.382	None
SVM	50.00%	61.79%	50.00%	50.00%	0.5303	0.382	None

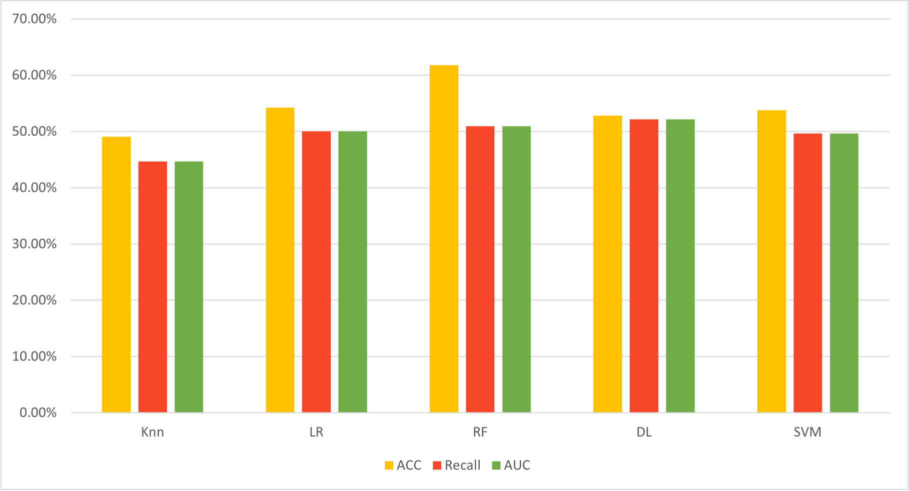


Figure A.7: *Openness trait recall measure and random forest as feature reduction is applied to the machine learning algorithms.*

Table A.7: *Openness trait scores based on recall fit and random forest as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.70%	49.06%	44.64%	44.64%	0.9999	0.5094	-0.1111
LR	57.94%	54.25%	50.02%	50.01%	0.9896	0.4575	0.0003
RF	51.78%	61.79%	50.94%	50.94%	0.5303	0.382	0.048
DL	56.09%	52.83%	**52.17%**	52.17%	0.9968	0.4716	0.0423
SVM	56.31%	53.77%	49.64%	49.63%	0.9928	0.4622	-0.0075

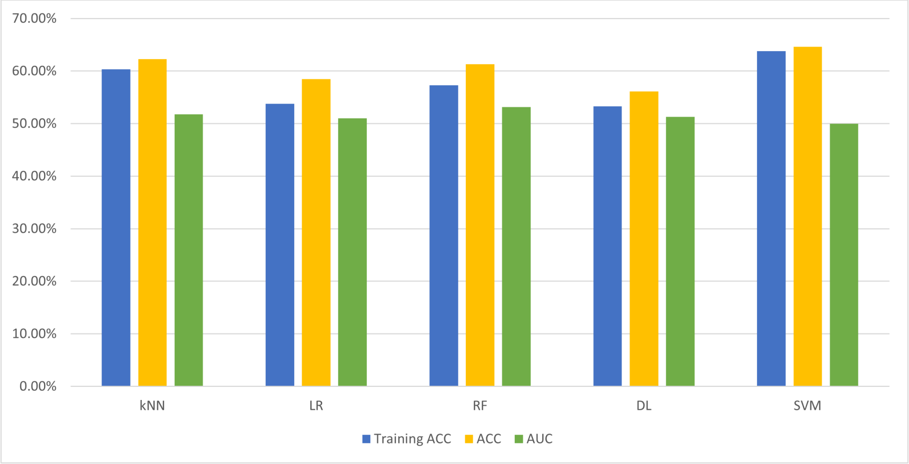


Figure A.8: *Conscientiousness trait accuracy measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.8: *Conscientiousness trait scores based on accuracy fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	60.31%	62.26%	51.80%	51.79%	0.7859	0.3773	0.0499
LR	53.77%	58.49%	50.99%	50.98%	0.9727	0.415	0.0221
RF	57.27%	61.32%	53.18%	53.17%	0.8591	0.3867	0.0743
DL	53.32%	56.13%	51.27%	51.27%	0.9956	0.4386	0.0259
SVM	63.80%	**64.62%**	50.00%	50.00%	0.5313	0.3537	None

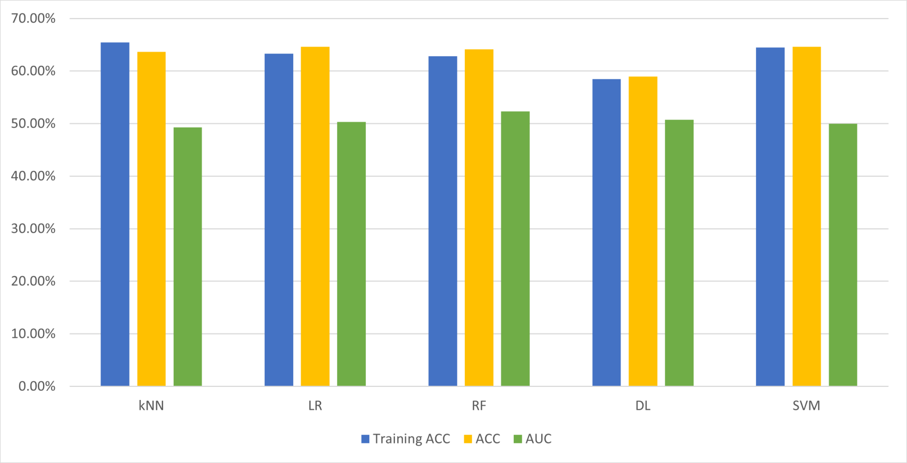


Figure A.9: *Conscientiousness trait accuracy measure and parameter tuning is applied to the machine learning algorithms.*

Table A.9: *Conscientiousness trait scores based on accuracy fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	65.42%	63.68%	49.27%	49.27%	0.6424	0.3632	-0.0722
LR	63.31%	**64.62%**	50.30%	50.30%	0.5313	0.3537	0.0298
RF	62.82%	64.15%	52.35%	52.35%	0.5878	0.3584	0.0786
DL	58.44%	58.96%	50.75%	50.74%	0.9626	0.4103	0.0173
SVM	64.48%	**64.62%**	50%	50%	0.5313	0.3537	None

Table A.10: *Conscientiousness trait scores based on recall fit and no parameter tuning is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.39%	62.26%	51.80%	51.79%	0.7859	0.3773	0.0499
LR	48.23%	58.49%	50.99%	50.98%	0.9727	0.415	0.0221
RF	51.12%	61.32%	**54.34%**	53.17%	0.8591	0.3867	0.0743
DL	50.73%	56.13%	51.27%	51.27%	0.9956	0.4386	0.0259
SVM	51.14%	64.62%	50.00%	53.13%	0.3537	0.5	None

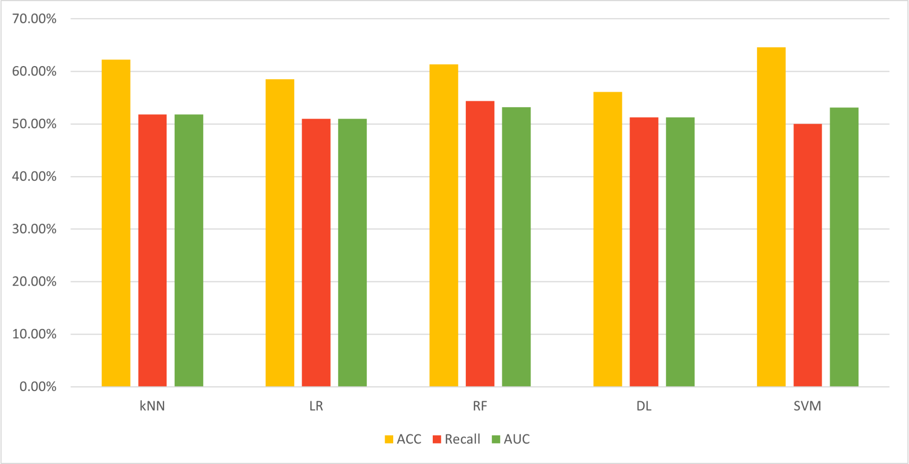


Figure A.10: *Conscientiousness trait recall measure and no parameter tuning is applied to the machine learning algorithms.*

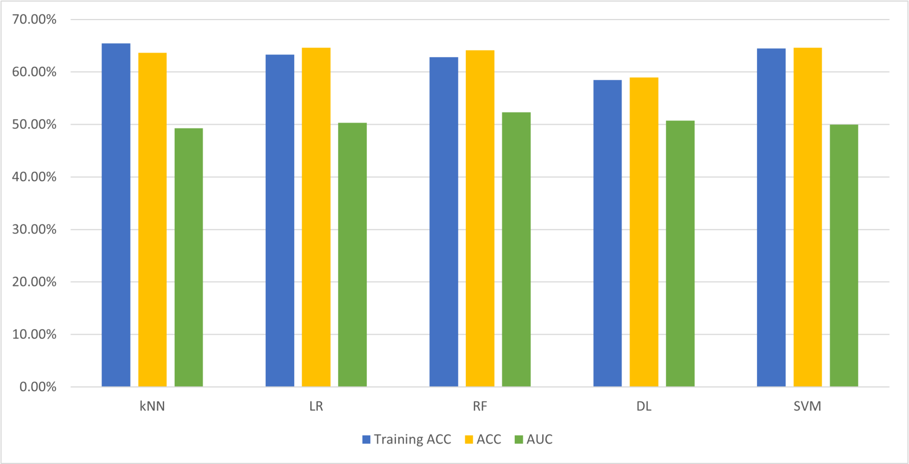


Figure A.11: *Conscientiousness trait recall measure and parameter tuning is applied to the machine learning algorithms.*

Table A.11: *Conscientiousness trait scores based on recall fit and parameter tuning is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	54.43%	61.32%	51.37%	51.36%	0.8591	0.3867	0.036
LR	51.96%	58.49%	51.29%	51.28%	0.9727	0.415	0.0284
RF	51.42%	63.21%	**51.92%**	51.92%	0.6942	0.3679	0.0591
DL	53.26%	55.19%	50.55%	50.54%	0.9981	0.4481	0.011
SVM	50.00%	64.62%	50.00%	50.00%	0.5313	0.3537	None

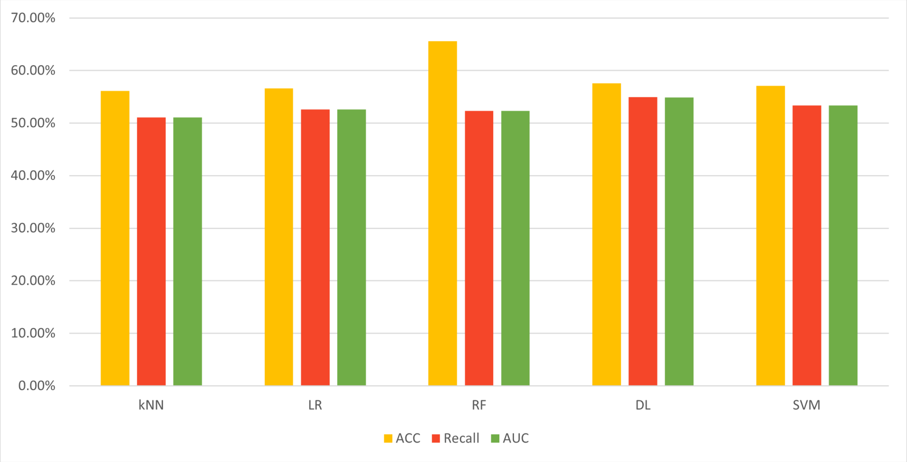


Figure A.12: *Conscientiousness trait recall measure and ANOVA as feature reduction is applied to the machine learning algorithms.*

Table A.12: *Conscientiousness trait scores based on recall fit and ANOVA as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	52.98%	56.13%	51.07%	51.07%	0.9999	0.4386	0.0207
LR	68.38%	56.60%	52.62%	52.61%	0.9998	0.4339	0.0501
RF	55.08%	65.57%	52.35%	52.34%	0.8317	0.3443	0.0644
DL	69.28%	57.55%	**54.92%**	54.91%	0.9996	0.4245	0.0929
SVM	67.44%	57.08%	53.37%	53.36%	0.9997	0.4292	0.0642

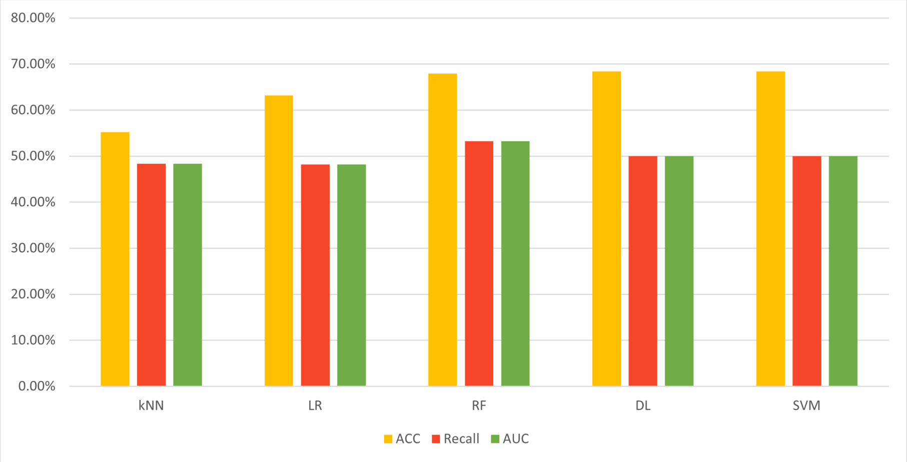


Figure A.13: *Conscientiousness trait recall measure and LASSO as feature reduction is applied to the machine learning algorithms.*

Table A.13: *Conscientiousness trait scores based on recall fit and LASSO as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.75%	55.19%	48.37%	48.37%	0.9999	0.4481	-0.0323
LR	53.14%	63.21%	48.21%	48.21%	0.9538	0.3679	-0.0555
RF	53.76%	67.92%	**53.27%**	53.26%	0.591	0.3207	0.1063
DL	49.94%	68.40%	50.00%	50.00%	0.533	0.316	None
SVM	50.00%	68.40%	50.00%	50.00%	0.533	0.316	None

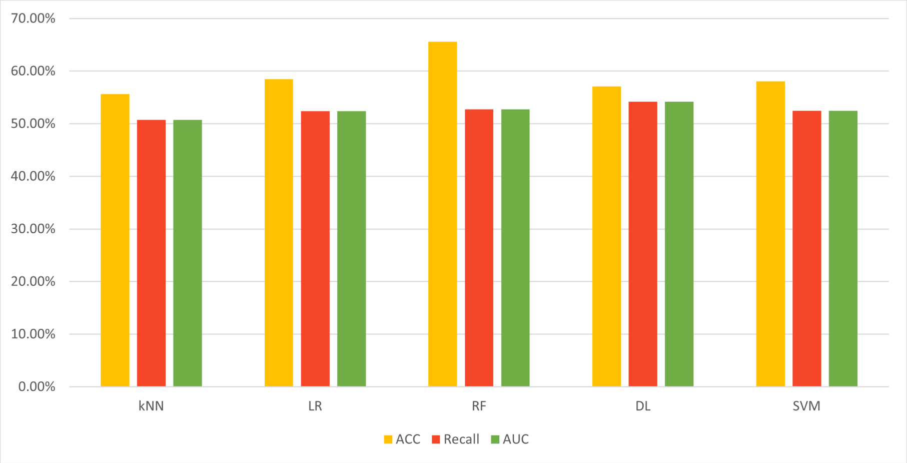


Figure A.14: *Conscientiousness trait recall measure and random forest as feature reduction is applied to the machine learning algorithms.*

Table A.14: *Conscientiousness trait scores based on recall fit and random forest as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.07%	55.66%	50.73%	50.72%	0.9999	0.4433	0.014
LR	61.77%	58.49%	52.39%	52.39%	0.999	0.415	0.0474
RF	54.03%	65.57%	52.75%	52.74%	0.8317	0.3443	0.0733
DL	60.76%	57.08%	**54.17%**	54.16%	0.9997	0.4292	0.0789
SVM	58.50%	58.02%	52.45%	52.44%	0.9994	0.4198	0.0481

Table A.15: *Extraversion trait scores based on accuracy fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	45.80%	52.83%	52.41%	52.41%	0.5279	0.4716	0.0486
LR	47.40%	**55.66%**	55.46%	55.46%	0.2248	0.4433	0.1094
RF	49.72%	53.77%	53.46%	53.46%	0.4187	0.4622	0.0696
DL	48.35%	50%	50.48%	50.48%	0.8144	0.5	0.0097
SVM	47.86%	50%	48.88%	48.87%	0.8144	0.5	-0.0244

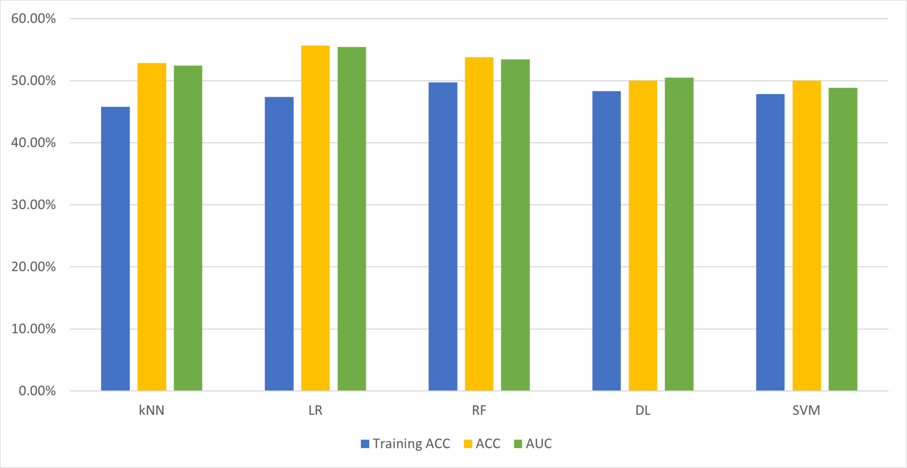


Figure A.15: *Extraversion trait accuracy measure and no parameter tuning is applied to the machine learning algorithms.*

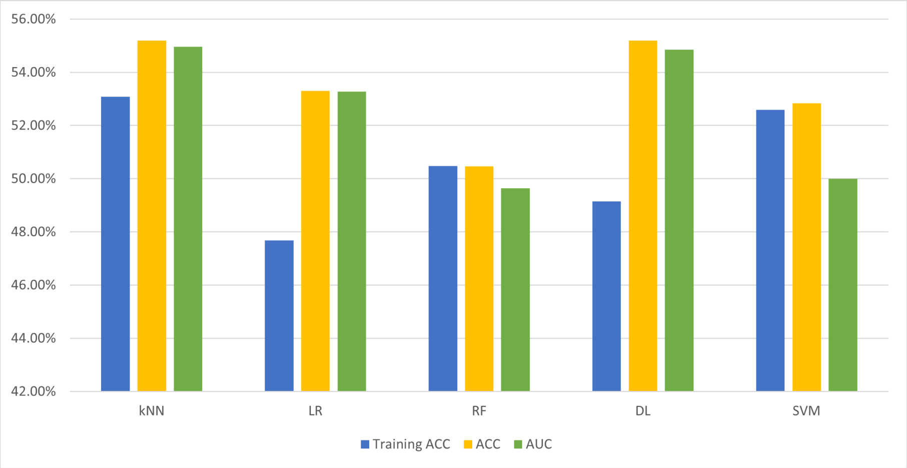


Figure A.16: *Extraversion trait accuracy measure and parameter tuning is applied to the machine learning algorithms.*

Table A.16: *Extraversion trait scores based on accuracy fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	53.09%	**55.19%**	54.96%	54.96%	0.2682	0.4481	0.0994
LR	47.68%	53.30%	53.29%	53.28%	0.4731	0.4669	0.0656
RF	50.48%	50.47%	49.64%	49.64%	0.7755	0.4952	-0.0074
DL	49.15%	**55.19%**	54.86%	54.85%	0.2682	0.4481	0.0977
SVM	52.59%	52.83%	50%	50%	0.5279	0.4716	None

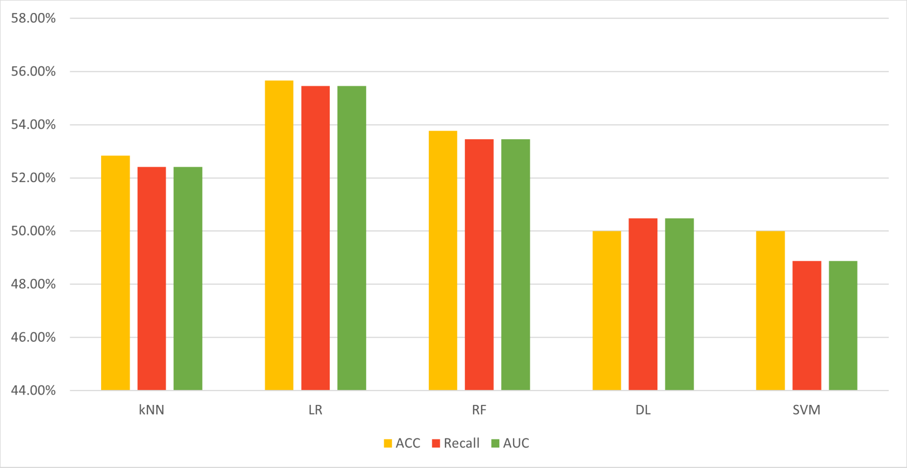


Figure A.17: *Extraversion trait recall measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.17: *Extraversion trait scores based on recall fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	46.58%	52.83%	52.41%	52.41%	0.5279	0.4716	0.0486
LR	47.95%	55.66%	**55.46%**	55.46%	0.2248	0.4433	0.1094
RF	51.13%	53.77%	53.46%	53.46%	0.4187	0.4622	0.0696
DL	48.84%	50.00%	50.48%	50.48%	0.8144	0.5	0.0097
SVM	46.20%	50.00%	48.88%	48.88%	0.8144	0.5	-0.0244

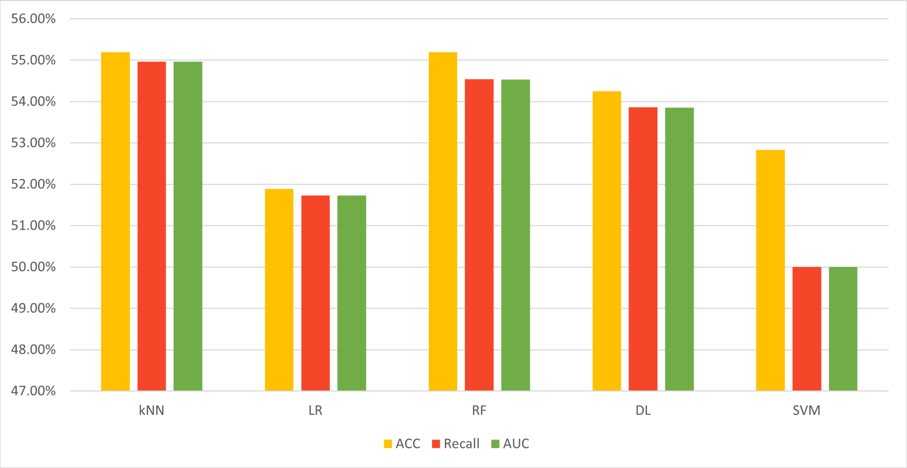


Figure A.18: *Extraversion trait recall measure and parameter tuning is applied to the machine learning algorithms.*

Table A.18: *Extraversion trait scores based on recall fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	52.92%	55.19%	**54.96%**	54.96%	0.2682	0.4481	0.0994
LR	48.64%	51.89%	51.73%	51.73%	0.6349	0.4811	0.0346
RF	50.03%	55.19%	54.54%	54.53%	0.2682	0.4481	0.0931
DL	49.90%	54.25%	53.86%	53.85%	0.3659	0.4575	0.0778
SVM	50.00%	52.83%	50.00%	50.00%	0.5279	0.4716	None

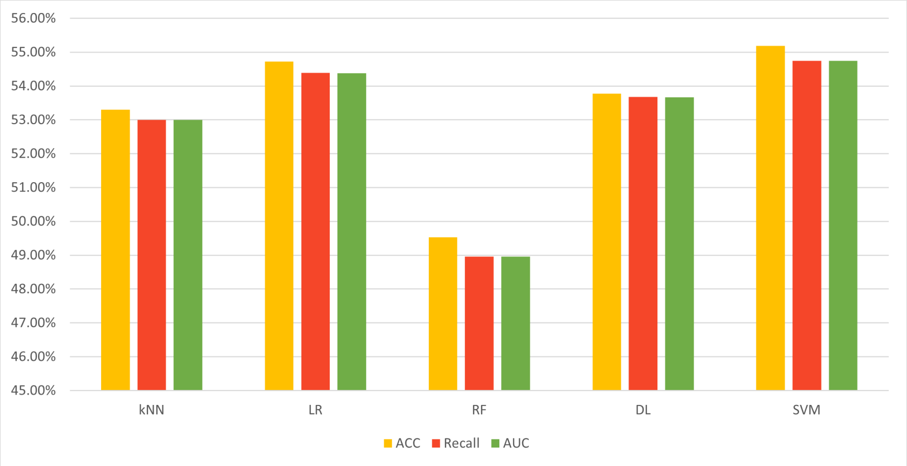


Figure A.19: *Extraversion trait recall measure and ANOVA as feature reduction is applied to the machine learning algorithms.*

Table A.19: *Extraversion trait scores based on recall fit and ANOVA as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	52.50%	53.30%	53.00%	53.00%	0.6354	0.4669	0.06
LR	70.72%	54.72%	54.39%	54.38%	0.4733	0.4528	0.0877
RF	59.94%	49.53%	48.96%	48.96%	0.9259	0.5047	-0.0209
DL	71.15%	53.77%	53.68%	53.67%	0.5826	0.4622	0.0733
SVM	70.43%	55.19%	**54.74%**	54.74%	0.4188	0.4481	0.0951

Table A.20: *Extraversion trait scores based on recall fit and LASSO as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	50.55%	50.47%	50.56%	50.55%	0.8792	0.4952	0.011
LR	51.02%	50.94%	48.01%	48.00%	0.8493	0.4905	-0.0548
RF	58.34%	55.66%	**54.61%**	54.61%	0.3658	0.4433	0.095
DL	50.00%	54.25%	50.00%	50.00%	0.5282	0.4575	None
SVM	48.62%	49.53%	46.78%	46.78%	0.9259	0.5047	-0.0836

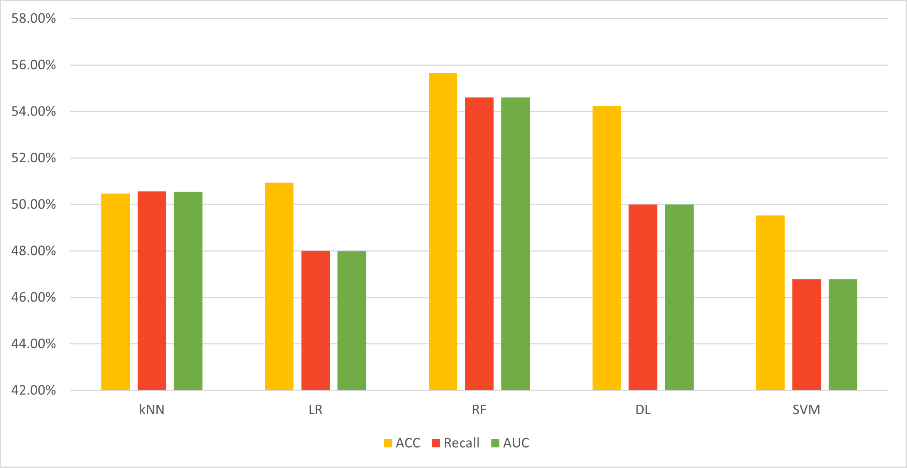


Figure A.20: *Extraversion trait recall measure and LASSO as feature reduction is applied to the machine learning algorithms.*

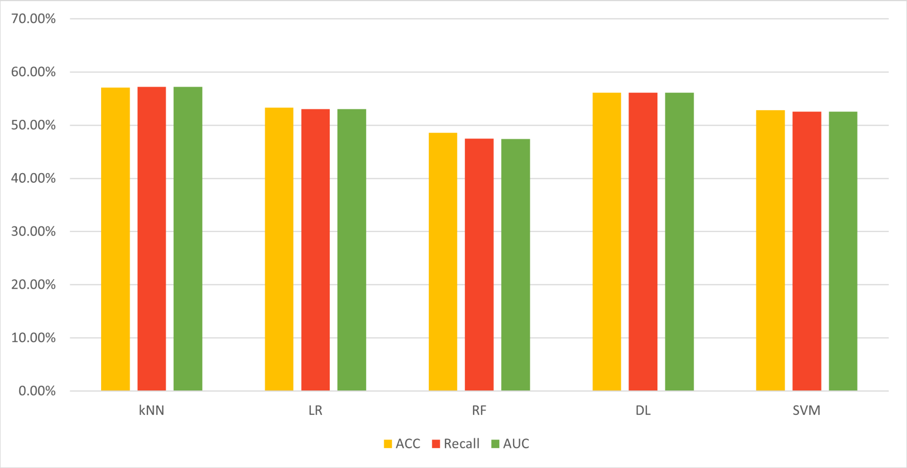


Figure A.21: *Extraversion trait recall measure and random forest as feature reduction is applied to the machine learning algorithms.*

Table A.21: *Extraversion trait scores based on recall fit and random forest as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.40%	57.08%	**57.21%**	57.20%	0.2244	0.4292	0.1436
LR	64.28%	53.30%	53.00%	53.00%	0.6354	0.4669	0.06
RF	57.51%	48.58%	47.45%	47.44%	0.9573	0.5141	-0.0527
DL	64.61%	56.13%	56.10%	56.09%	0.3153	0.4386	0.1214
SVM	64.00%	52.83%	52.57%	52.56%	0.6857	0.4716	0.0512

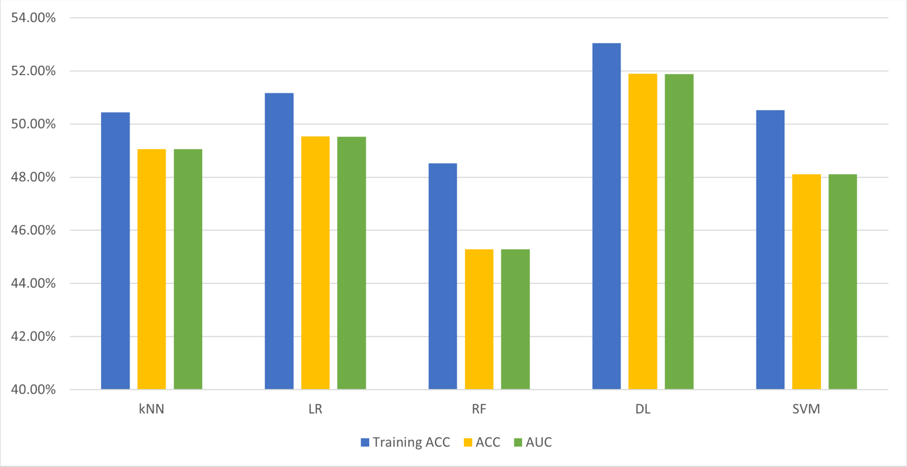


Figure A.22: *Agreeableness trait accuracy measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.22: *Agreeableness trait scores based on accuracy fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	50.44%	49.06%	49.06%	49.05%	0.6343	0.5094	-0.0192
LR	51.17%	49.53%	49.53%	49.52%	0.5815	0.5047	-0.0095
RF	48.52%	45.28%	45.28%	45.28%	0.9254	0.5471	-0.0943
DL	53.04%	**51.89%**	51.89%	51.88%	0.3153	0.4811	0.0377
SVM	50.53%	48.11%	48.11%	48.11%	0.7317	0.5188	-0.0377

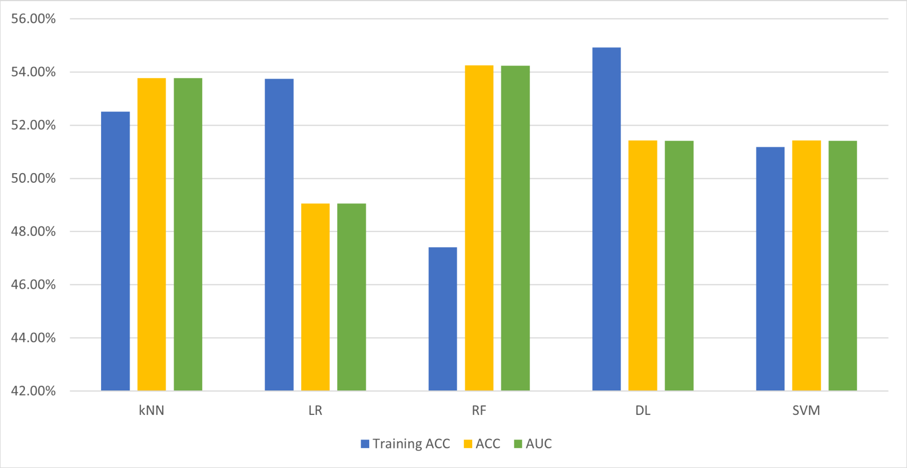


Figure A.23: *Agreeableness trait accuracy measure and parameter tuning is applied to the machine learning algorithms.*

Table A.23: *Agreeableness trait scores based on accuracy fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	52.51%	53.77%	53.77%	53.77%	0.1514	0.4622	0.0763
LR	53.74%	49.06%	49.06%	49.05%	0.6343	0.5094	-0.0189
RF	47.41%	**54.25%**	54.25%	54.24%	0.1214	0.4575	0.0849
DL	54.92%	51.42%	51.42%	51.41%	0.3656	0.4858	0.0283
SVM	51.18%	51.42%	51.42%	51.41%	0.3656	0.4858	0.0284

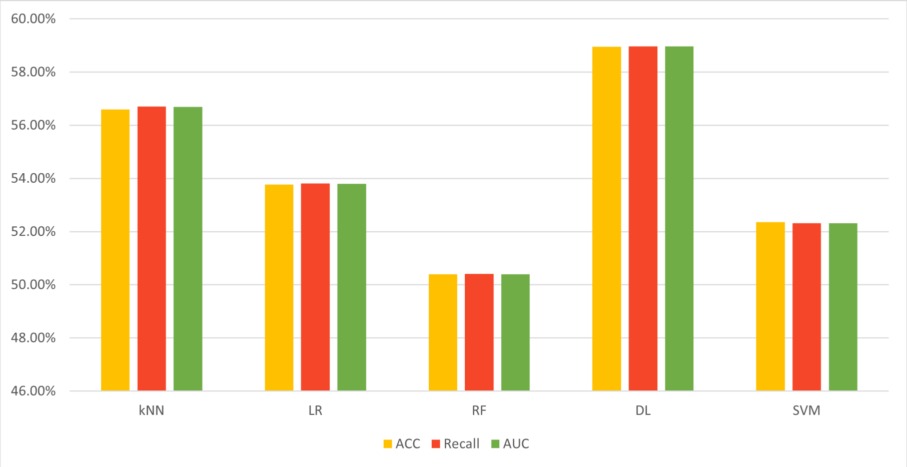


Figure A.24: *Agreeableness trait recall measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.24: *Agreeableness trait scores based on recall fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	48.50%	56.60%	56.70%	56.69%	0.0428	0.4339	0.1366
LR	54.02%	53.77%	53.81%	53.80%	0.1859	0.4622	0.0762
RF	48.00%	50.40%	50.41%	50.40%	0.5274	0.4952	0.0082
DL	55.28%	58.96%	**58.97%**	58.97%	0.0079	0.4103	0.1794
SVM	51.50%	52.36%	52.32%	52.32%	0.3154	0.4764	0.0465

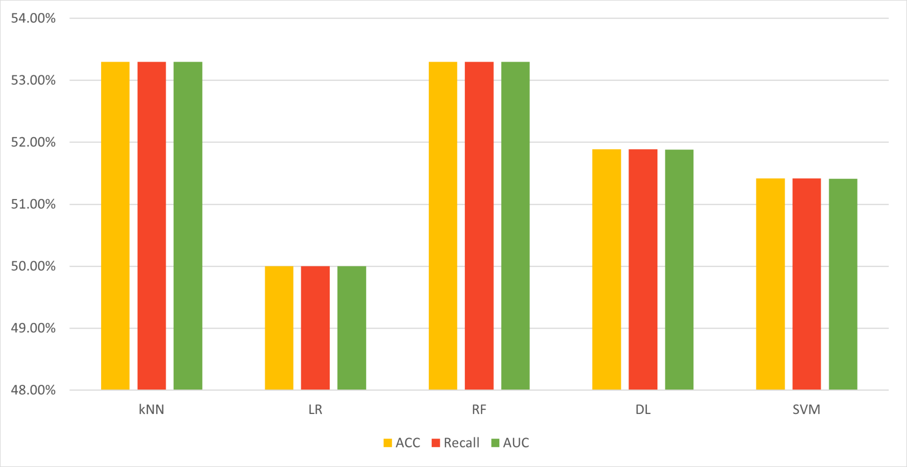


Figure A.25: *Agreeableness trait recall measure and parameter tuning is applied to the machine learning algorithms.*

Table A.25: *Agreeableness trait scores based on recall fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	51.92%	53.30%	**53.30%**	53.30%	0.1859	0.4669	0.0679
LR	52.34%	50.00%	50.00%	50.00%	0.5273	0.5	0.5
RF	46.92%	53.30%	**53.30%**	53.30%	0.1859	0.4669	0.0661
DL	52.07%	51.89%	51.89%	51.88%	0.3153	0.4811	0.0377
SVM	52.09%	51.42%	51.42%	51.41%	0.3656	0.4858	0.0284

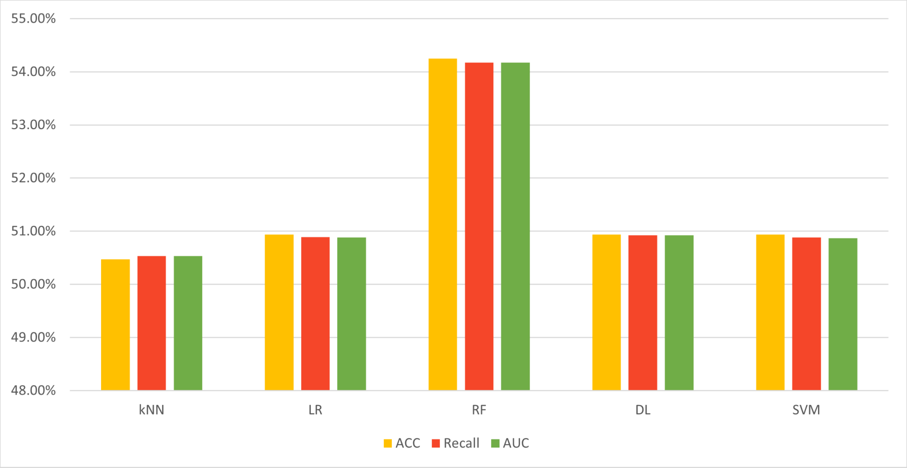


Figure A.26: *Agreeableness trait recall measure and ANOVA as feature reduction is applied to the machine learning algorithms.*

Table A.26: *Agreeableness trait scores based on recall fit and ANOVA as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	47.45%	50.47%	50.53%	50.53%	0.5274	0.4952	0.0107
LR	66.36%	50.94%	50.89%	50.88%	0.4727	0.4905	0.0178
RF	58.39%	54.25%	**54.17%**	54.17%	0.1514	0.4575	0.0844
DL	68.27%	50.94%	50.92%	50.92%	0.4727	0.4905	0.0184
SVM	64.48%	50.94%	50.88%	50.87%	0.4727	0.4905	0.0177

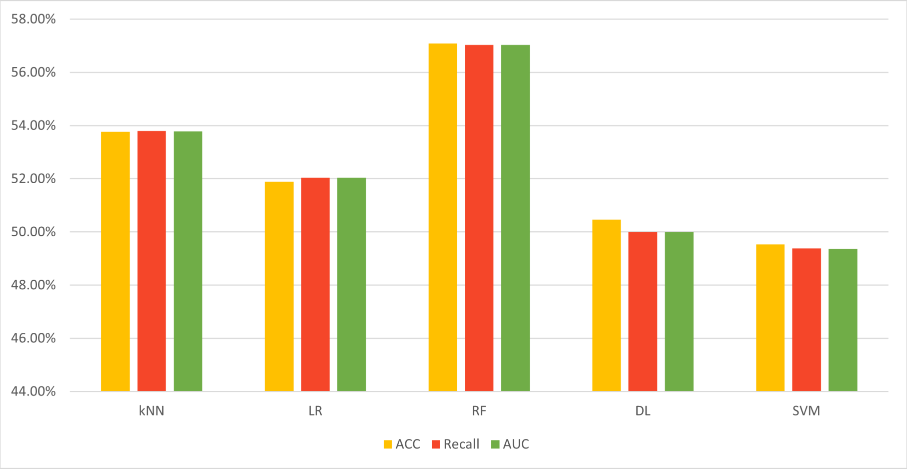


Figure A.27: *Agreeableness trait recall measure and LASSO as feature reduction is applied to the machine learning algorithms.*

Table A.27: *Agreeableness trait scores based on recall fit and LASSO as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	57.28%	53.77%	53.79%	53.78%	0.1859	0.4622	0.0757
LR	47.52%	51.89%	52.04%	52.04%	0.3657	0.4811	0.0432
RF	54.58%	57.08%	**57.03%**	57.03%	0.0316	0.4292	0.1412
DL	50.00%	50.47%	50.00%	50.00%	0.5274	0.4952	None
SVM	49.58%	49.53%	49.38%	49.37%	0.6343	0.5047	-0.0131

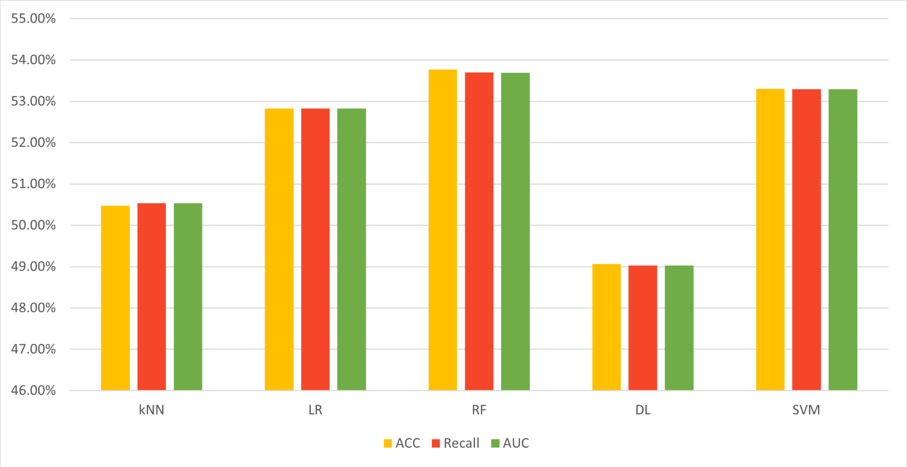


Figure A.28: *Agreeableness trait recall measure and random forest as feature reduction is applied to the machine learning algorithms.*

Table A.28: *Agreeableness trait scores based on recall fit and random forest as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	46.80%	50.47%	50.53%	50.53%	0.5274	0.4952	0.0107
LR	58.42%	52.83%	52.83%	52.83%	0.2683	0.4716	0.0566
RF	53.18%	53.77%	**53.70%**	53.69%	0.1859	0.4622	0.0749
DL	57.97%	49.06%	49.03%	49.03%	0.6846	0.5094	-0.0193
SVM	59.34%	53.30%	53.29%	53.29%	0.225	0.4669	0.0658

Table A.29: *Neuroticism trait scores based on accuracy fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	54.67%	50%	49.75%	49.75%	0.7753	0.5	-0.0049
LR	48.85%	45.75%	45.61%	45.61%	0.9768	0.5424	-0.0878
RF	47.67%	**50.47%**	50.21%	50.20%	0.7321	0.4952	0.0041
DL	50.50%	45.75%	45.61%	45.61%	0.9768	0.5424	-0.0878
SVM	45.09%	45.75%	45.34%	45.34%	0.9768	0.5424	-0.0943

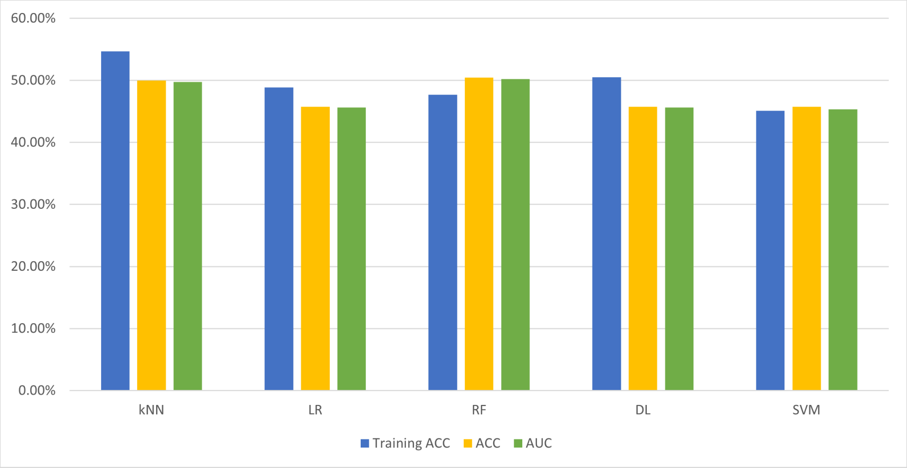


Figure A.29: *Neuroticism trait accuracy measure and no parameter tuning is applied to the machine learning algorithms.*

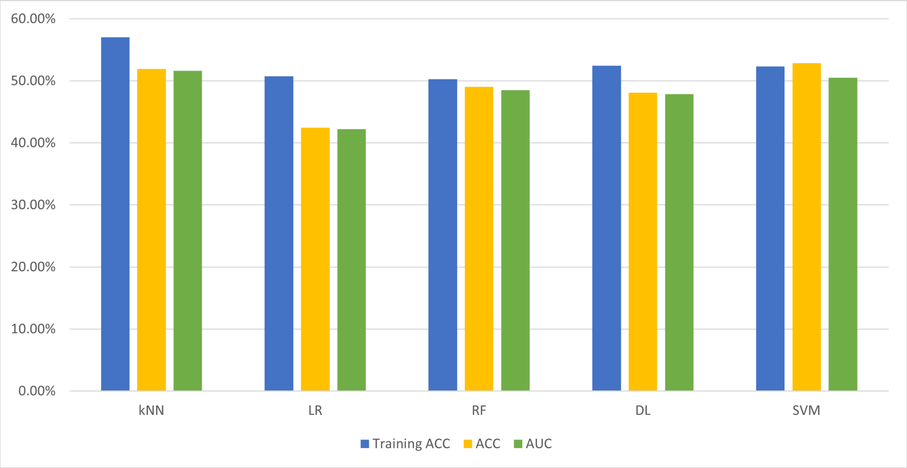


Figure A.30: *Neuroticism trait accuracy measure and parameter tuning is applied to the machine learning algorithms.*

Table A.30: *Neuroticism trait scores based on accuracy fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	57.04%	51.89%	51.65%	51.64%	0.4811	0.4811	0.033
LR	50.76%	42.45%	42.24%	42.23%	0.5754	0.5754	-0.1556
RF	50.26%	49.06%	48.50%	48.49%	0.5094	0.5094	-0.0308
DL	52.43%	48.11%	47.86%	47.86%	0.5188	0.5188	-0.0429
SVM	52.34%	**52.83%**	50.54%	50.53%	0.4716	0.4716	0.0456

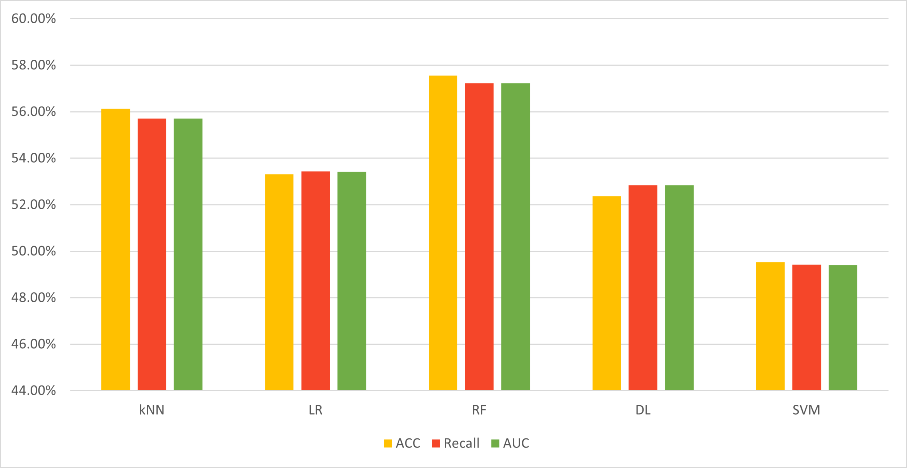


Figure A.31: *Neuroticism trait recall measure and no parameter tuning is applied to the machine learning algorithms.*

Table A.31: *Neuroticism trait scores based on recall fit and no parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	55.01%	56.13%	55.70%	55.70%	0.268	0.4386	0.1145
LR	49.66%	53.30%	53.43%	53.42%	0.5824	0.4669	0.0683
RF	46.77%	57.55%	**57.23%**	57.23%	0.1507	0.4245	0.1448
DL	46.07%	52.36%	52.84%	52.83%	0.6855	0.4764	0.057
SVM	48.94%	49.53%	49.42%	49.41%	0.9045	0.5047	-0.0116

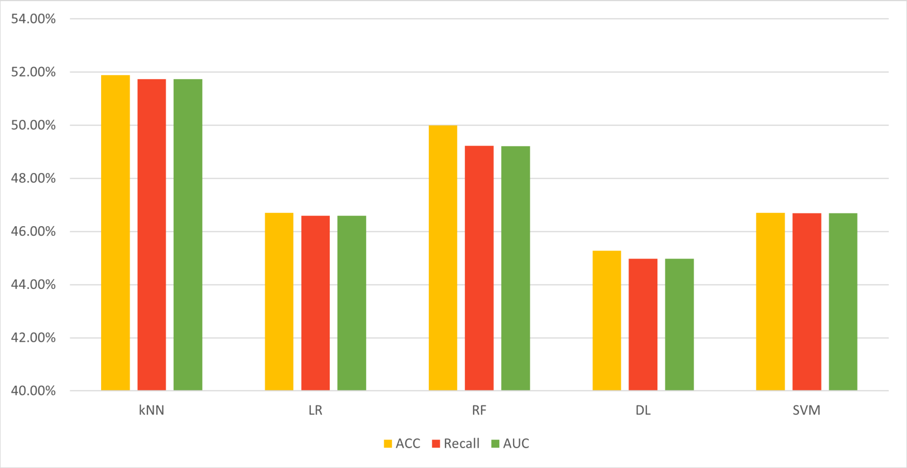


Figure A.32: *Neuroticism trait recall measure and parameter tuning is applied to the machine learning algorithms.*

Table A.32: *Neuroticism trait scores based on recall fit and parameter tuning is applied. Highest accuracy score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	55.25%	51.89%	51.73%	51.73%	0.582	0.4811	0.0347
LR	51.95%	46.70%	46.60%	46.60%	0.9571	0.533	-0.0679
RF	48.33%	50.00%	**49.22%**	49.21%	0.7753	0.5	-0.0165
DL	52.36%	45.28%	44.98%	44.98%	0.9834	0.5471	-0.1009
SVM	51.07%	46.70%	46.69%	46.69%	0.9571	0.533	-0.0661

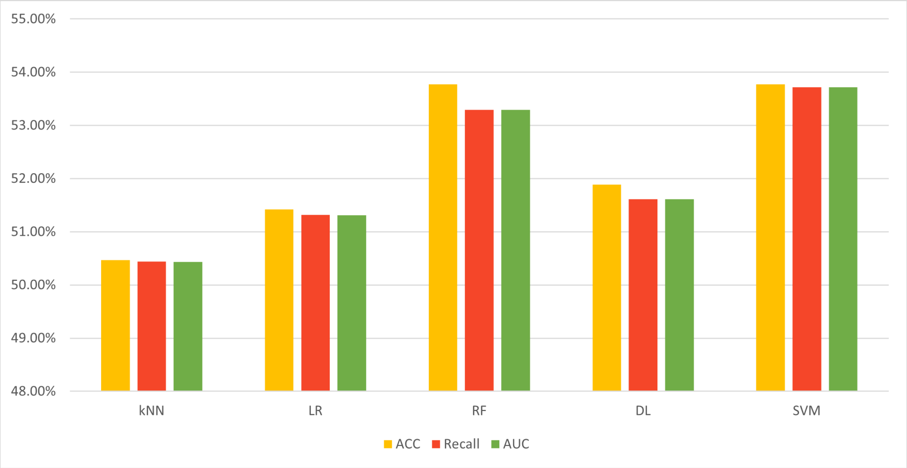


Figure A.33: *Neuroticism trait recall measure and ANOVA as feature reduction is applied to the machine learning algorithms.*

Table A.33: *Neuroticism trait scores based on recall fit and ANOVA as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	53.73%	50.47%	50.44%	50.43%	0.8492	0.4952	0.0087
LR	68.23%	51.42%	51.32%	51.31%	0.7758	0.4858	0.0262
RF	51.71%	53.77%	53.29%	53.29%	0.5281	0.4622	0.0662
DL	69.22%	51.89%	51.61%	51.61%	0.7326	0.4811	0.0322
SVM	66.82%	53.77%	**53.72%**	53.72%	0.5281	0.4622	0.0742

Table A.34: *Neuroticism trait scores based on recall fit and LASSO as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	53.35%	46.70%	46.93%	46.92%	0.9834	0.533	-0.0613
LR	48.59%	46.70%	45.50%	45.49%	0.9834	0.533	-0.0944
RF	53.20%	46.70%	46.29%	46.28%	0.9834	0.533	-0.0744
DL	50.48%	54.72%	**51.02%**	51.02%	0.4188	0.4528	0.1052
SVM	49.77%	53.77%	50.00%	50.00%	0.5281	0.4622	None

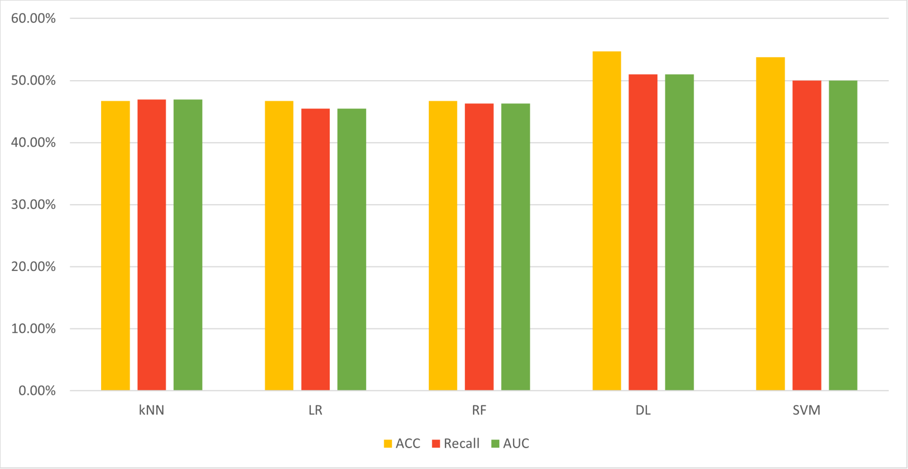


Figure A.34: *Neuroticism trait recall measure and LASSO as feature reduction is applied to the machine learning algorithms.*

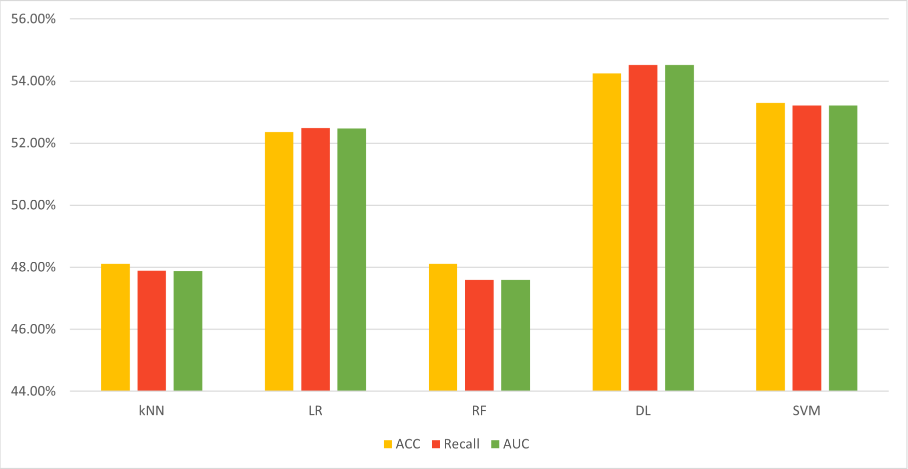


Figure A.35: *Neuroticism trait recall measure and random forest as feature reduction is applied to the machine learning algorithms.*

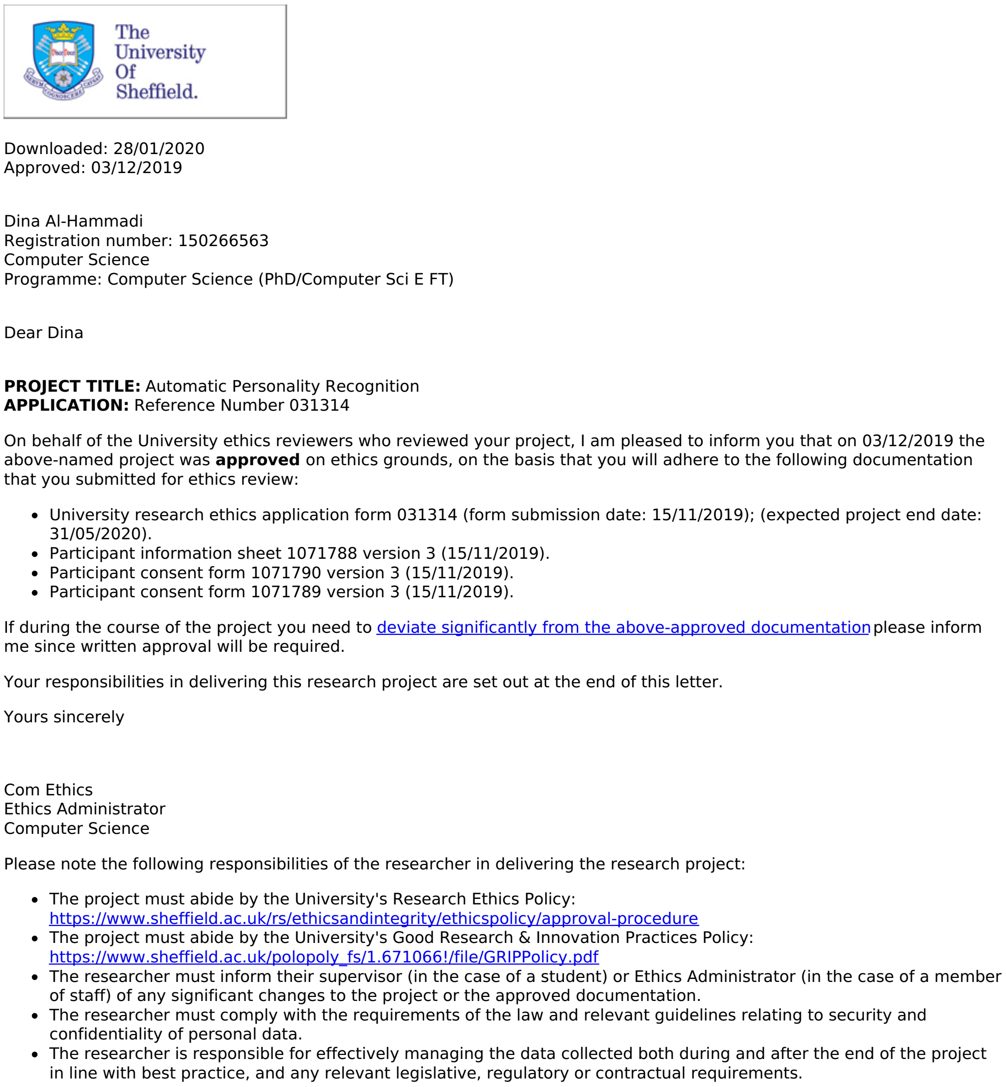
Table A.35: *Neuroticism trait scores based on recall fit and random forest as feature reduction is applied. Highest recall score is in bold.*

Classifier	Training ACC	ACC	Recall	AUC	P value	Error Rate	MCC
kNN	53.73%	48.11%	47.89%	47.88%	0.9572	0.5188	-0.0421
LR	53.73%	52.36%	52.48%	52.47%	0.6855	0.4764	0.0494
RF	52.87%	48.11%	47.60%	47.60%	0.9572	0.5188	0.476
DL	54.60%	54.25%	**54.52%**	54.52%	0.4732	0.4575	0.0903
SVM	53.52%	53.30%	53.21%	53.21%	0.5824	0.4669	0.0641

Appendix B: Data Collection Forms for Personality Traits Corpus

This appendix lists the official documents used for the data collection and the ethics application form approved by the University of Sheffield.

- The University of Sheffield's Ethics Approval Letter.
- UREC Information Sheet for Target Participant.
- UREC Information Sheet for Acquaintance Participant.
- UREC Consent Form for Target Participant.
- UREC Consent Form for Acquaintance Participant.
- BFI-44 Questionnaire for Target Participant.
- BFI-44 Questionnaire for Acquaintance Participant.
- Interview Questions for Target Participant.



The University Of Sheffield.

Downloaded: 28/01/2020
Approved: 03/12/2019

Dina Al-Hammadi
Registration number: 150266563
Computer Science
Programme: Computer Science (PhD/Computer Sci E FT)

Dear Dina

PROJECT TITLE: Automatic Personality Recognition
APPLICATION: Reference Number 031314

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 03/12/2019 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 031314 (form submission date: 15/11/2019); (expected project end date: 31/05/2020).
- Participant information sheet 1071788 version 3 (15/11/2019).
- Participant consent form 1071790 version 3 (15/11/2019).
- Participant consent form 1071789 version 3 (15/11/2019).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Com Ethics
Ethics Administrator
Computer Science

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/approval-procedure
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

The University Of Sheffield.

January 1st, 2020

Participant Information Sheet (Target Participant)

1. Research Project Title:

Automatic Personality Recognition

2. Invitation paragraph

You are being asked to participate in a research study designed to build a data set of audio recordings and personality ratings. How people's personalities are recognized and perceived through verbal and non-verbal cues. Which features are important for personality recognition and perception.

3. What is the project's purpose?

The aim of this project is to produce and make publicly available a "New Speaker Personality Date Set (Working Title)".

4. Why have I been chosen?

Your participation is voluntary. You can choose to participate or not.

5. Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time* without any negative consequences. You do not have to give a reason. If you wish to withdraw from the research, please contact me at this email: **dal-hammadi1@sheffield.ac.uk**

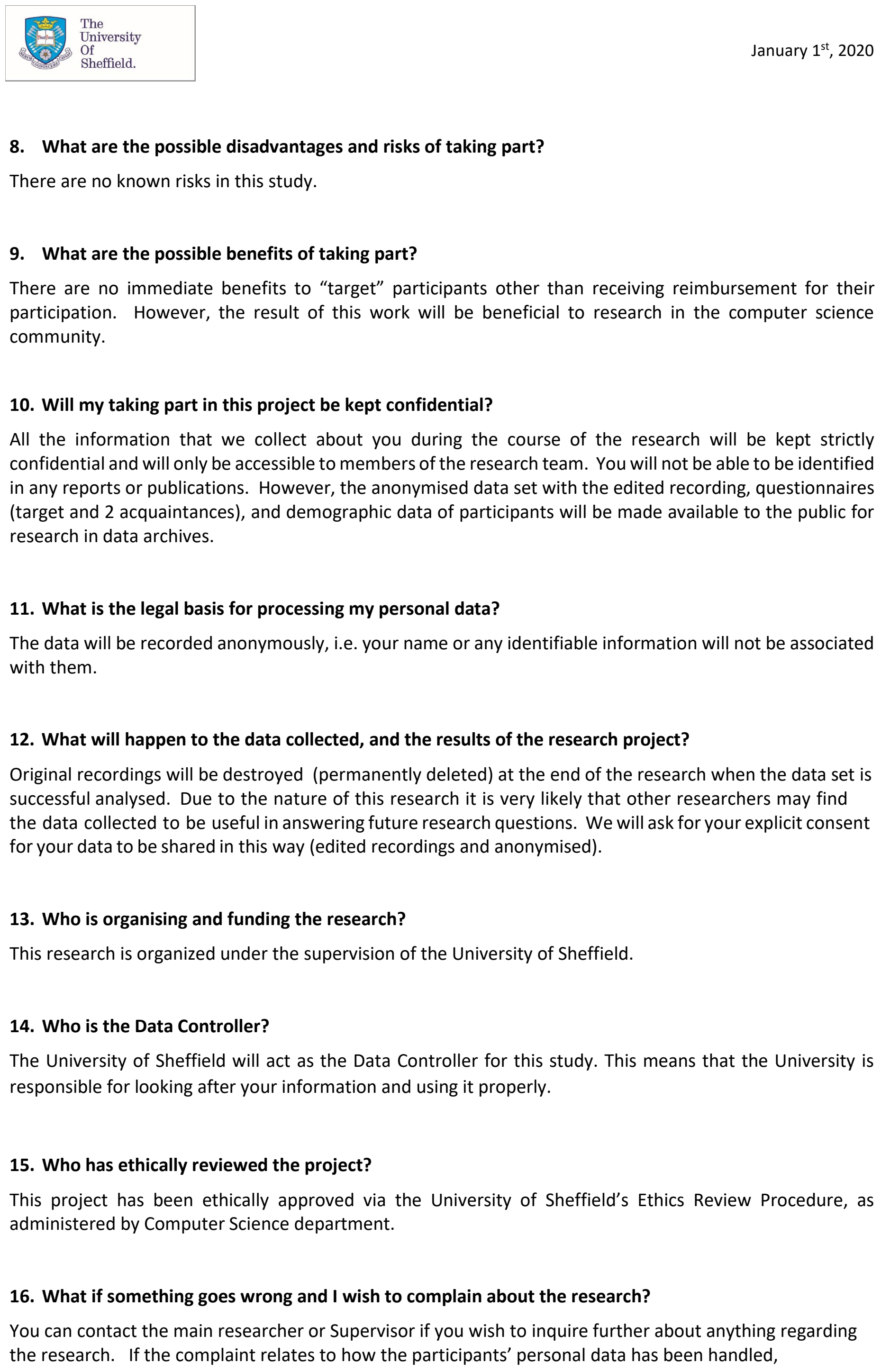
*Please note that there is a point at which it will not be possible for your data to be withdrawn from the research (once data have been anonymised and included within a large dataset). Please be aware that after **April 30st, 2020**, your data cannot be removed from the study beyond this point.

6. What will happen to me if I take part? What do I have to do?

- You (Target participant) will be interviewed, and the interview will be audio recorded. You will get a copy of the questions before sitting for the interview. The interview will last a maximum of 10 minutes (Later edited to 3-5 minutes). No identifiable information will be collected. Only demographic data about you will be collected. You will be asked to rate yourself using the Big Five Inventory (44 items) questionnaire.
- You (Target participant) must provide at least 2 acquaintances whom you have know for at least 6 months.
- Acquaintance participants will be asked to provide demographic data about themselves. And they will be asked to rate you (the target participant) using the Big Five Inventory (44 items) questionnaire.

7. Will I be recorded, and how will the recorded media be used?

The audio recordings of your interview made during this research will be used for analysis and for illustration in conference presentations and lectures. The final result which is an anonymised data set with full or partial recordings will be made available in public data archives for research purposes indefinitely.



January 1st, 2020

8. What are the possible disadvantages and risks of taking part?

There are no known risks in this study.

9. What are the possible benefits of taking part?

There are no immediate benefits to "target" participants other than receiving reimbursement for their participation. However, the result of this work will be beneficial to research in the computer science community.

10. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications. However, the anonymised data set with the edited recording, questionnaires (target and 2 acquaintances), and demographic data of participants will be made available to the public for research in data archives.

11. What is the legal basis for processing my personal data?

The data will be recorded anonymously, i.e. your name or any identifiable information will not be associated with them.

12. What will happen to the data collected, and the results of the research project?

Original recordings will be destroyed (permanently deleted) at the end of the research when the data set is successful analysed. Due to the nature of this research it is very likely that other researchers may find the data collected to be useful in answering future research questions. We will ask for your explicit consent for your data to be shared in this way (edited recordings and anonymised).

13. Who is organising and funding the research?

This research is organized under the supervision of the University of Sheffield.

14. Who is the Data Controller?

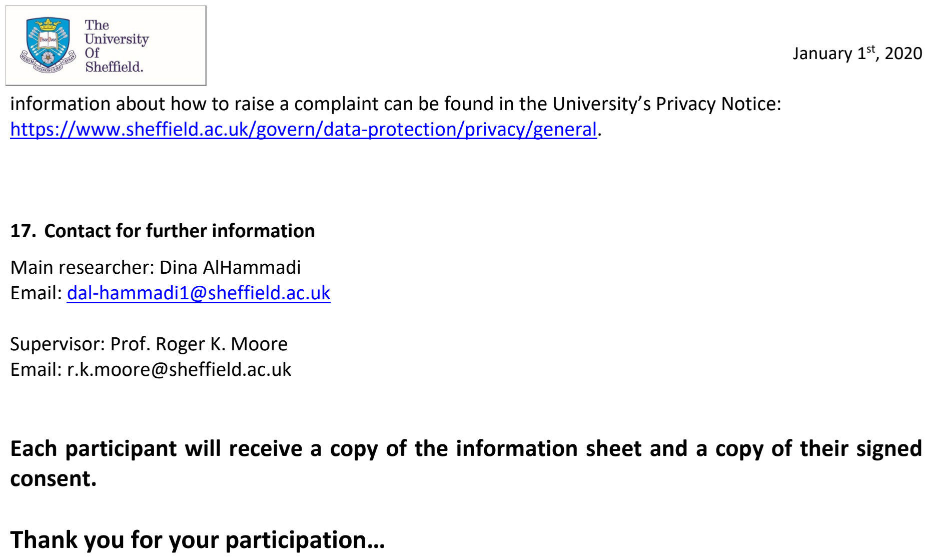
The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly.

15. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by Computer Science department.

16. What if something goes wrong and I wish to complain about the research?

You can contact the main researcher or Supervisor if you wish to inquire further about anything regarding the research. If the complaint relates to how the participants' personal data has been handled,



January 1st, 2020

information about how to raise a complaint can be found in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

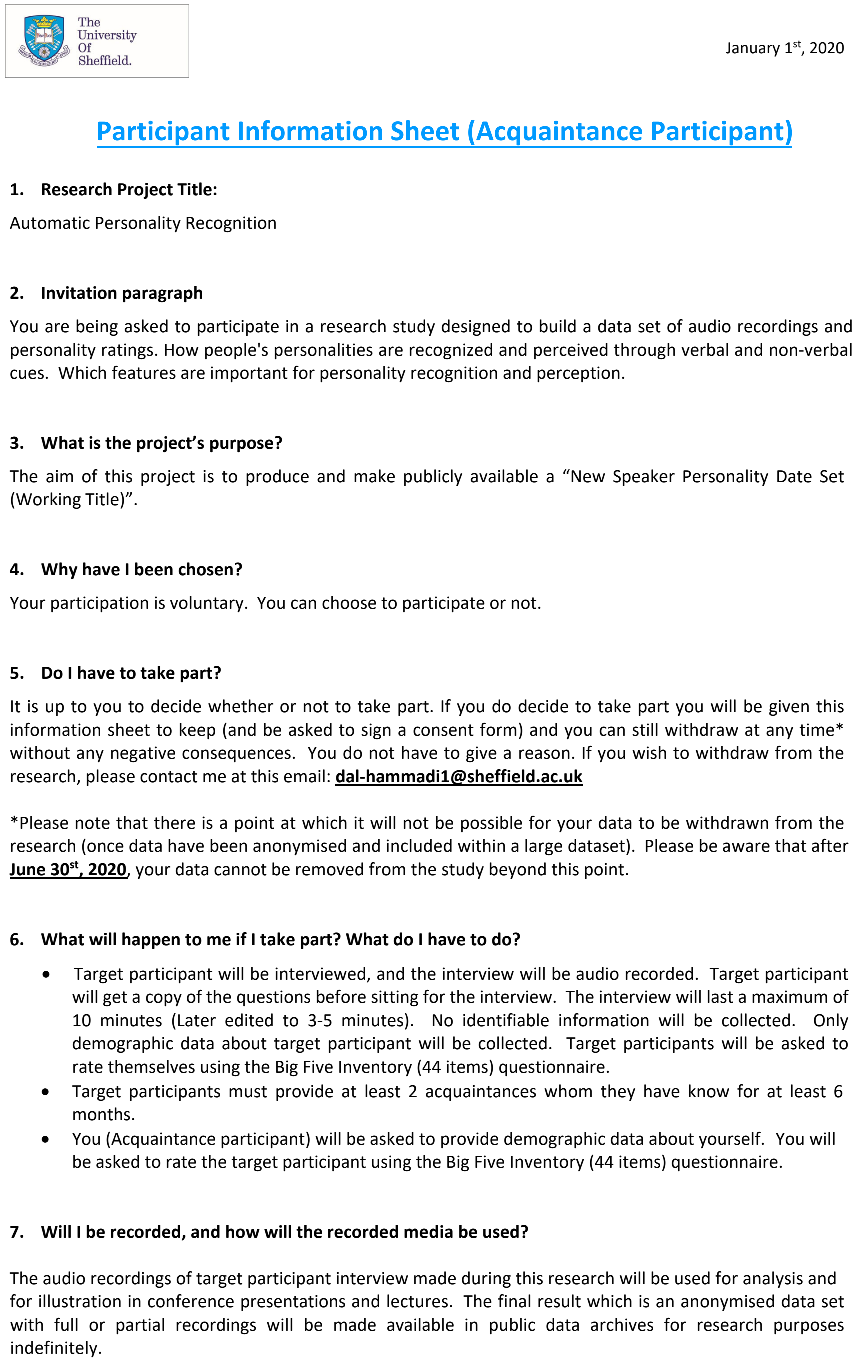
17. Contact for further information

Main researcher: Dina AlHammadi
Email: dal-hammadi1@sheffield.ac.uk

Supervisor: Prof. Roger K. Moore
Email: r.k.moore@sheffield.ac.uk

Each participant will receive a copy of the information sheet and a copy of their signed consent.

Thank you for your participation…



The University Of Sheffield.

January 1st, 2020

Participant Information Sheet (Acquaintance Participant)

1. Research Project Title:

Automatic Personality Recognition

2. Invitation paragraph

You are being asked to participate in a research study designed to build a data set of audio recordings and personality ratings. How people's personalities are recognized and perceived through verbal and non-verbal cues. Which features are important for personality recognition and perception.

3. What is the project's purpose?

The aim of this project is to produce and make publicly available a "New Speaker Personality Date Set (Working Title)".

4. Why have I been chosen?

Your participation is voluntary. You can choose to participate or not.

5. Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time* without any negative consequences. You do not have to give a reason. If you wish to withdraw from the research, please contact me at this email: **dal-hammadi1@sheffield.ac.uk**

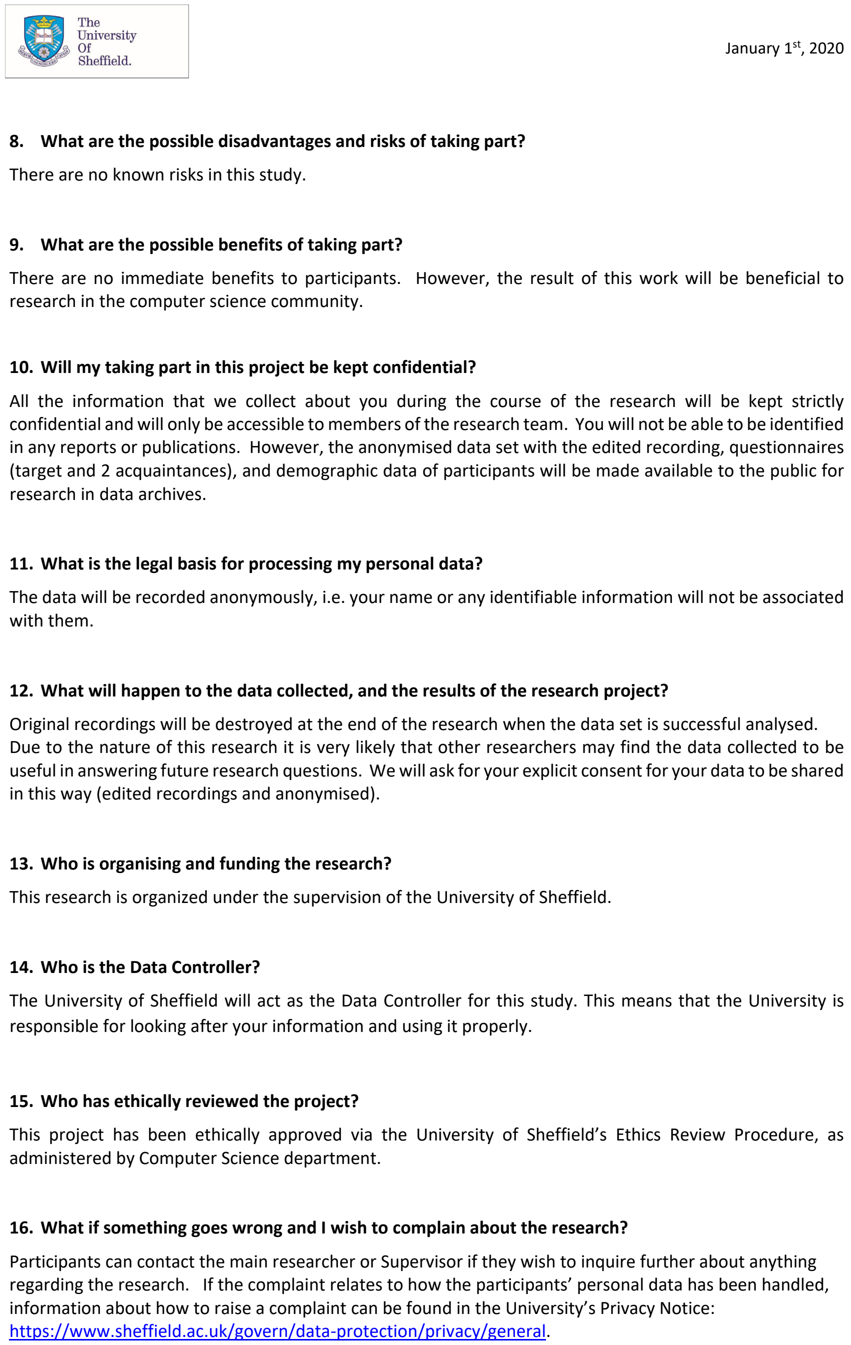
*Please note that there is a point at which it will not be possible for your data to be withdrawn from the research (once data have been anonymised and included within a large dataset). Please be aware that after **June 30st, 2020**, your data cannot be removed from the study beyond this point.

6. What will happen to me if I take part? What do I have to do?

- Target participant will be interviewed, and the interview will be audio recorded. Target participant will get a copy of the questions before sitting for the interview. The interview will last a maximum of 10 minutes (Later edited to 3-5 minutes). No identifiable information will be collected. Only demographic data about target participant will be collected. Target participants will be asked to rate themselves using the Big Five Inventory (44 items) questionnaire.
- Target participants must provide at least 2 acquaintances whom they have know for at least 6 months.
- You (Acquaintance participant) will be asked to provide demographic data about yourself. You will be asked to rate the target participant using the Big Five Inventory (44 items) questionnaire.

7. Will I be recorded, and how will the recorded media be used?

The audio recordings of target participant interview made during this research will be used for analysis and for illustration in conference presentations and lectures. The final result which is an anonymised data set with full or partial recordings will be made available in public data archives for research purposes indefinitely.



The
University
Of
Sheffield.

January 1st, 2020

8. What are the possible disadvantages and risks of taking part?

There are no known risks in this study.

9. What are the possible benefits of taking part?

There are no immediate benefits to participants. However, the result of this work will be beneficial to research in the computer science community.

10. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications. However, the anonymised data set with the edited recording, questionnaires (target and 2 acquaintances), and demographic data of participants will be made available to the public for research in data archives.

11. What is the legal basis for processing my personal data?

The data will be recorded anonymously, i.e. your name or any identifiable information will not be associated with them.

12. What will happen to the data collected, and the results of the research project?

Original recordings will be destroyed at the end of the research when the data set is successful analysed. Due to the nature of this research it is very likely that other researchers may find the data collected to be useful in answering future research questions. We will ask for your explicit consent for your data to be shared in this way (edited recordings and anonymised).

13. Who is organising and funding the research?

This research is organized under the supervision of the University of Sheffield.

14. Who is the Data Controller?

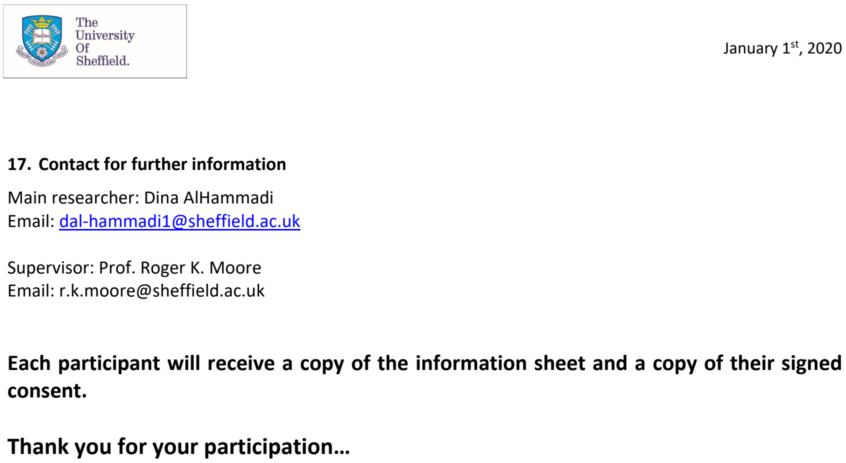
The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly.

15. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by Computer Science department.

16. What if something goes wrong and I wish to complain about the research?

Participants can contact the main researcher or Supervisor if they wish to inquire further about anything regarding the research. If the complaint relates to how the participants' personal data has been handled, information about how to raise a complaint can be found in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.



The
University
Of
Sheffield.

January 1st, 2020

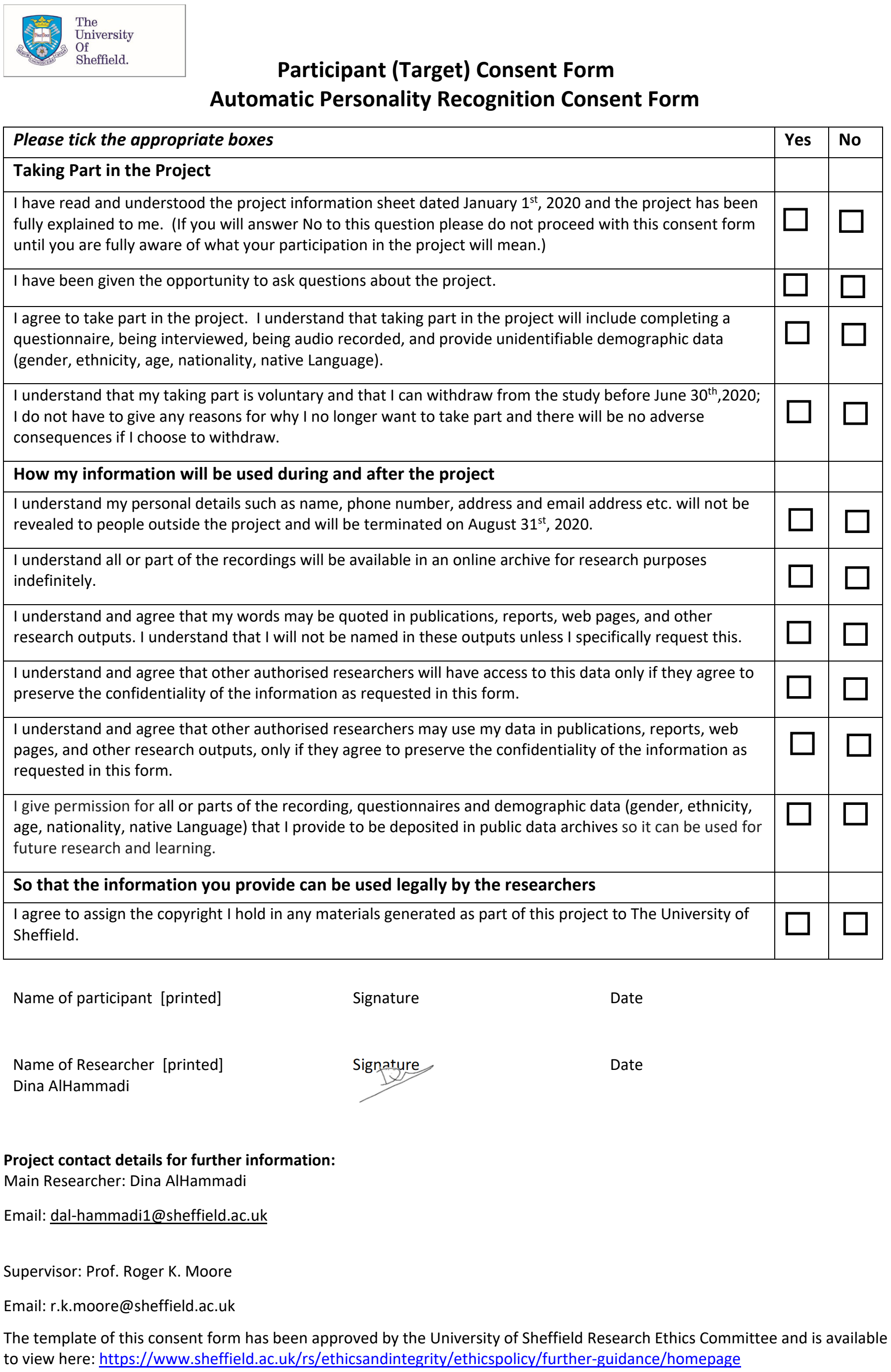
17. Contact for further information

Main researcher: Dina AlHammadi
Email: dal-hammadi1@sheffield.ac.uk

Supervisor: Prof. Roger K. Moore
Email: r.k.moore@sheffield.ac.uk

Each participant will receive a copy of the information sheet and a copy of their signed consent.

Thank you for your participation…



The University Of Sheffield.

Participant (Target) Consent Form
Automatic Personality Recognition Consent Form

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet dated January 1st, 2020 and the project has been fully explained to me. (If you will answer No to this question please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)	□	□
I have been given the opportunity to ask questions about the project.	□	□
I agree to take part in the project. I understand that taking part in the project will include completing a questionnaire, being interviewed, being audio recorded, and provide unidentifiable demographic data (gender, ethnicity, age, nationality, native Language).	□	□
I understand that my taking part is voluntary and that I can withdraw from the study before June 30th,2020; I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.	□	□
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project and will be terminated on August 31st, 2020.	□	□
I understand all or part of the recordings will be available in an online archive for research purposes indefinitely.	□	□
I understand and agree that my words may be quoted in publications, reports, web pages, and other research outputs. I understand that I will not be named in these outputs unless I specifically request this.	□	□
I understand and agree that other authorised researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand and agree that other authorised researchers may use my data in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I give permission for all or parts of the recording, questionnaires and demographic data (gender, ethnicity, age, nationality, native Language) that I provide to be deposited in public data archives so it can be used for future research and learning.	□	□
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	□	□

Name of participant [printed] Signature Date

Name of Researcher [printed] Signature Date
Dina AlHammadi

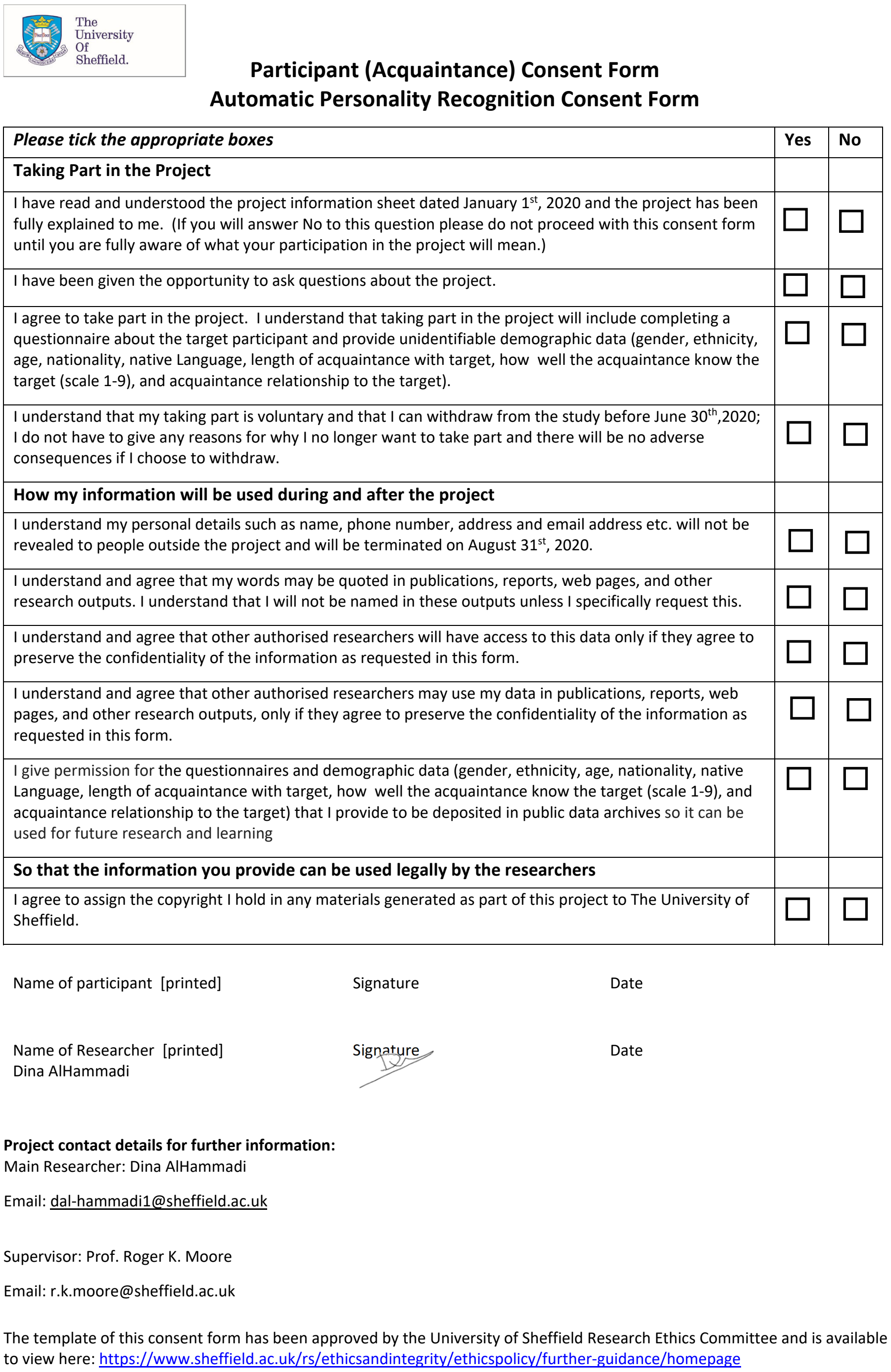
Project contact details for further information:
Main Researcher: Dina AlHammadi

Email: dal-hammadi1@sheffield.ac.uk

Supervisor: Prof. Roger K. Moore

Email: r.k.moore@sheffield.ac.uk

The template of this consent form has been approved by the University of Sheffield Research Ethics Committee and is available to view here: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/further-guidance/homepage



The University Of Sheffield.

Participant (Acquaintance) Consent Form
Automatic Personality Recognition Consent Form

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet dated January 1st, 2020 and the project has been fully explained to me. (If you will answer No to this question please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)	□	□
I have been given the opportunity to ask questions about the project.	□	□
I agree to take part in the project. I understand that taking part in the project will include completing a questionnaire about the target participant and provide unidentifiable demographic data (gender, ethnicity, age, nationality, native Language, length of acquaintance with target, how well the acquaintance know the target (scale 1-9), and acquaintance relationship to the target).	□	□
I understand that my taking part is voluntary and that I can withdraw from the study before June 30th,2020; I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.	□	□
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project and will be terminated on August 31st, 2020.	□	□
I understand and agree that my words may be quoted in publications, reports, web pages, and other research outputs. I understand that I will not be named in these outputs unless I specifically request this.	□	□
I understand and agree that other authorised researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand and agree that other authorised researchers may use my data in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I give permission for the questionnaires and demographic data (gender, ethnicity, age, nationality, native Language, length of acquaintance with target, how well the acquaintance know the target (scale 1-9), and acquaintance relationship to the target) that I provide to be deposited in public data archives so it can be used for future research and learning	□	□
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	□	□

Name of participant [printed] Signature Date

Name of Researcher [printed] Signature Date
Dina AlHammadi

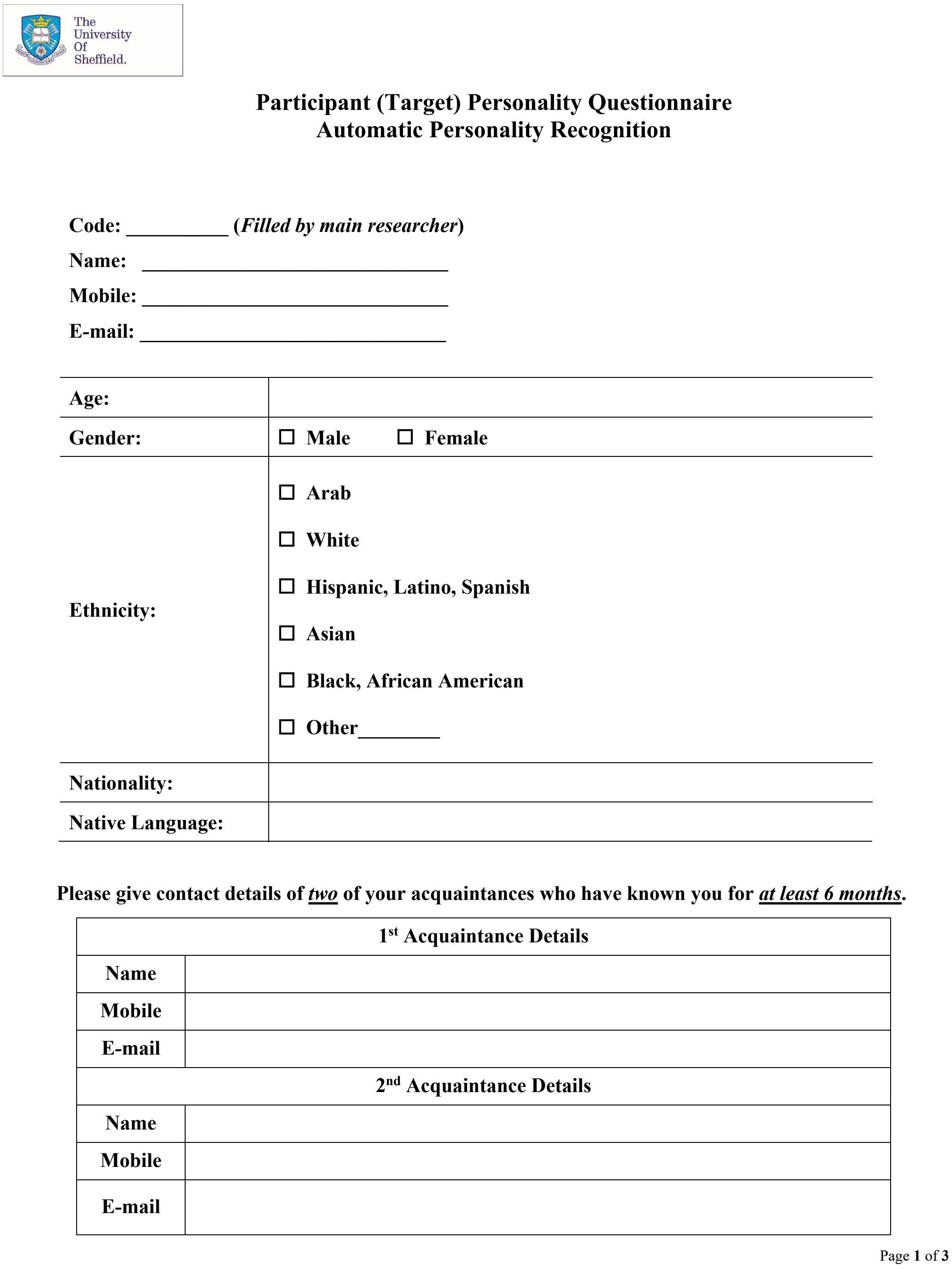
Project contact details for further information:
Main Researcher: Dina AlHammadi

Email: dal-hammadi1@sheffield.ac.uk

Supervisor: Prof. Roger K. Moore

Email: r.k.moore@sheffield.ac.uk

The template of this consent form has been approved by the University of Sheffield Research Ethics Committee and is available to view here: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/further-guidance/homepage



The University Of Sheffield.

Participant (Target) Personality Questionnaire
Automatic Personality Recognition

Code: __________ (*Filled by main researcher*)

Name: ______________________________

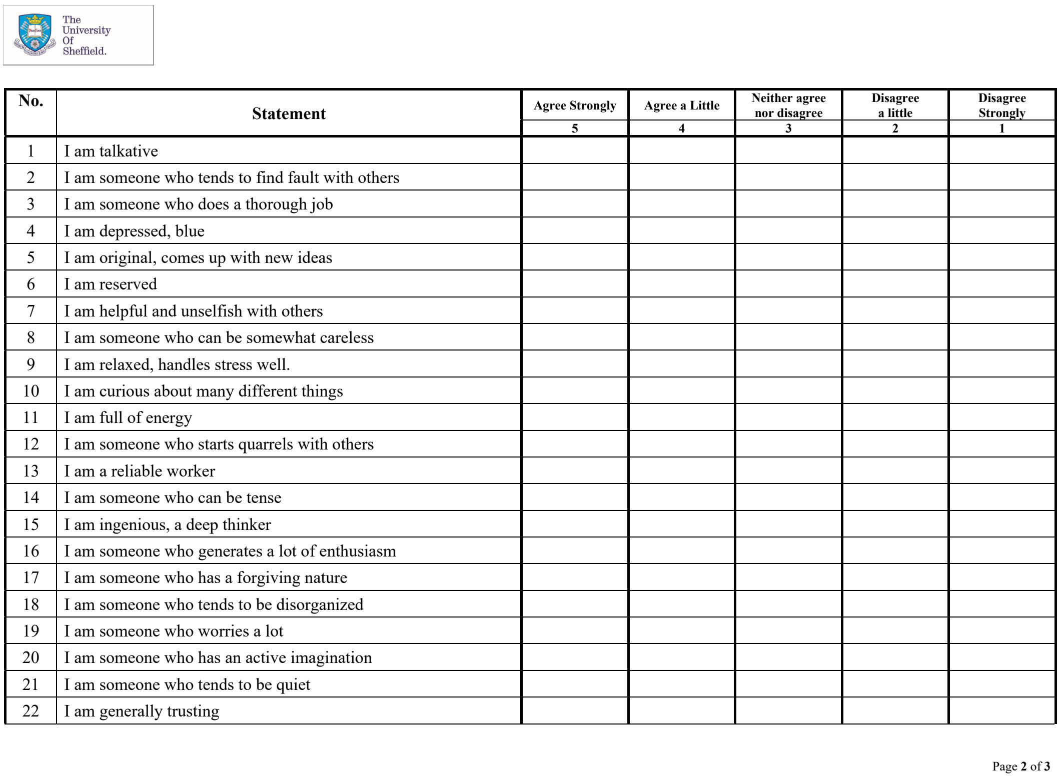
Mobile: ______________________________

E-mail: ______________________________

Age:	
Gender:	□ **Male** □ **Female**
Ethnicity:	□ **Arab** □ **White** □ **Hispanic, Latino, Spanish** □ **Asian** □ **Black, African American** □ **Other________**
Nationality:	
Native Language:	

Please give contact details of *<u>two</u>* of your acquaintances who have known you for *<u>at least 6 months</u>*.

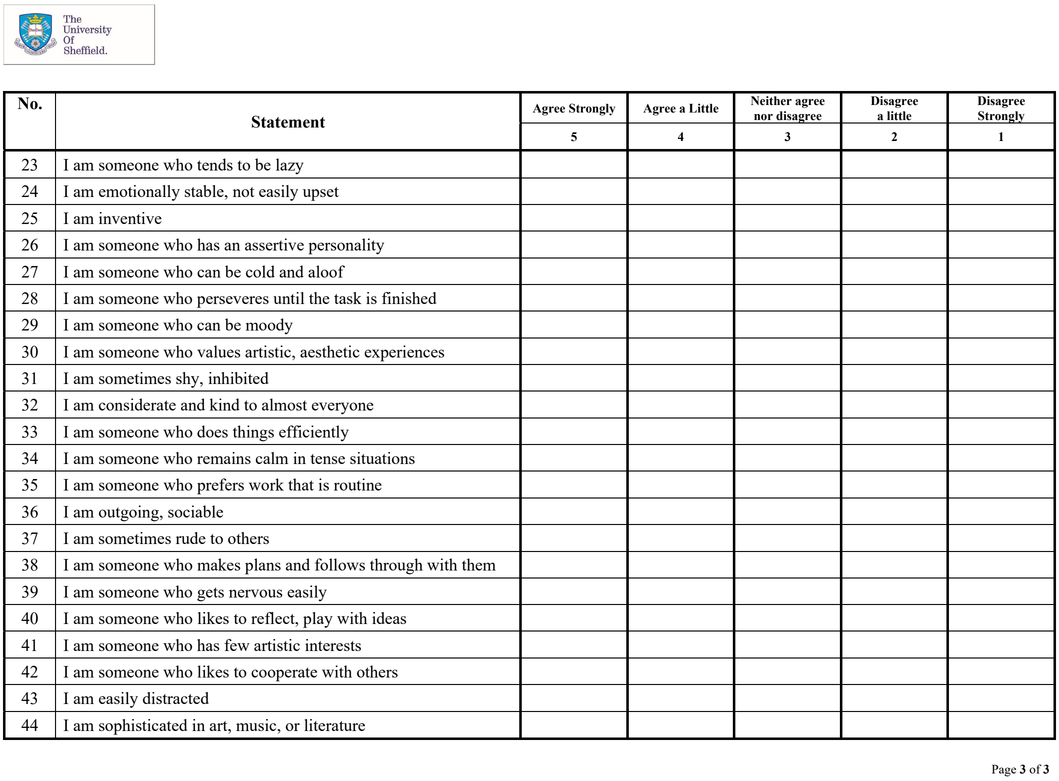
1st Acquaintance Details	
Name	
Mobile	
E-mail	
2nd Acquaintance Details	
Name	
Mobile	
E-mail	



The
University
Of
Sheffield.

No.	Statement	Agree Strongly	Agree a Little	Neither agree nor disagree	Disagree a little	Disagree Strongly
		5	4	3	2	1
1	I am talkative					
2	I am someone who tends to find fault with others					
3	I am someone who does a thorough job					
4	I am depressed, blue					
5	I am original, comes up with new ideas					
6	I am reserved					
7	I am helpful and unselfish with others					
8	I am someone who can be somewhat careless					
9	I am relaxed, handles stress well.					
10	I am curious about many different things					
11	I am full of energy					
12	I am someone who starts quarrels with others					
13	I am a reliable worker					
14	I am someone who can be tense					
15	I am ingenious, a deep thinker					
16	I am someone who generates a lot of enthusiasm					
17	I am someone who has a forgiving nature					
18	I am someone who tends to be disorganized					
19	I am someone who worries a lot					
20	I am someone who has an active imagination					
21	I am someone who tends to be quiet					
22	I am generally trusting					

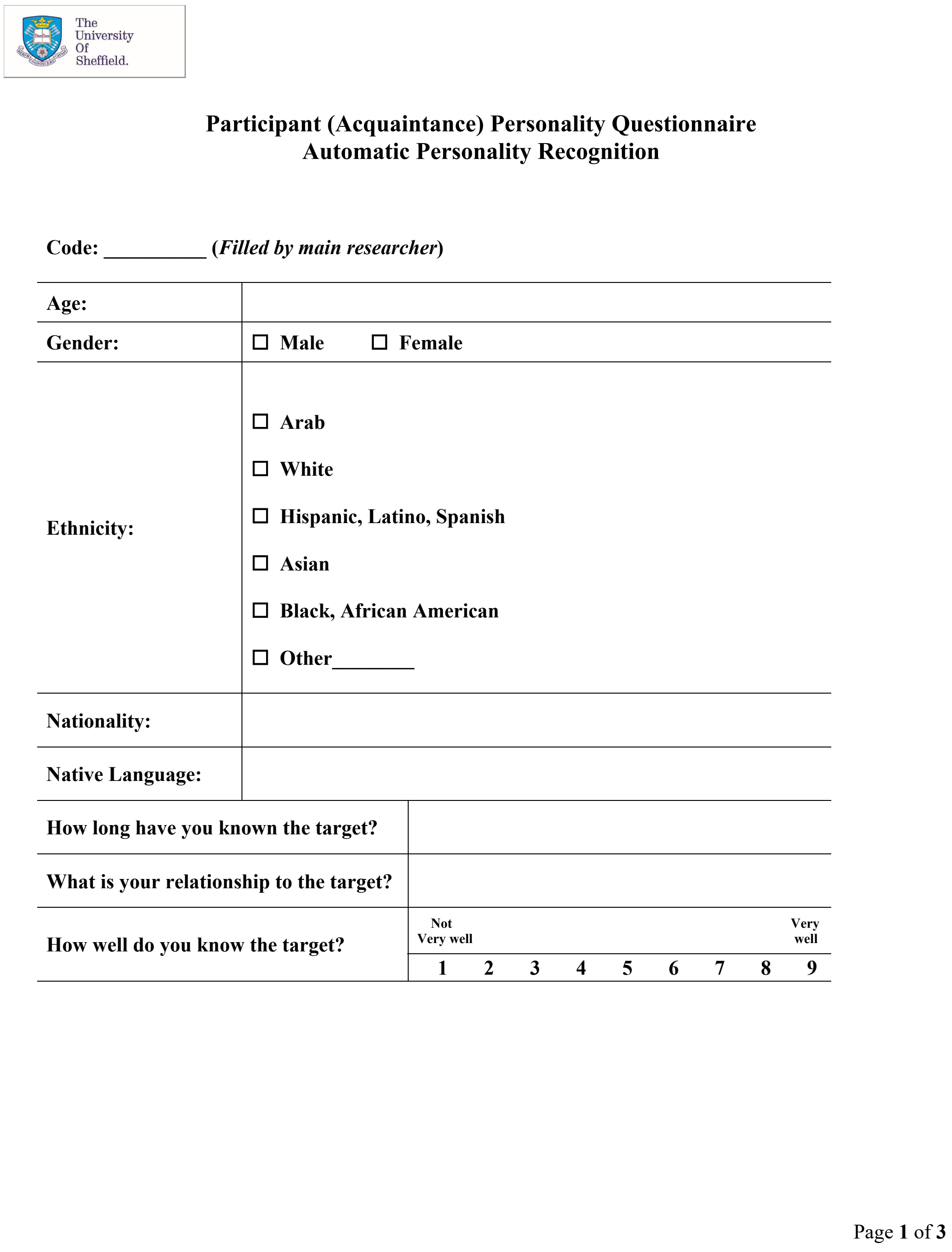
Page **2** of **3**



The University Of Sheffield.

No.	Statement	Agree Strongly	Agree a Little	Neither agree nor disagree	Disagree a little	Disagree Strongly
		5	4	3	2	1
23	I am someone who tends to be lazy					
24	I am emotionally stable, not easily upset					
25	I am inventive					
26	I am someone who has an assertive personality					
27	I am someone who can be cold and aloof					
28	I am someone who perseveres until the task is finished					
29	I am someone who can be moody					
30	I am someone who values artistic, aesthetic experiences					
31	I am sometimes shy, inhibited					
32	I am considerate and kind to almost everyone					
33	I am someone who does things efficiently					
34	I am someone who remains calm in tense situations					
35	I am someone who prefers work that is routine					
36	I am outgoing, sociable					
37	I am sometimes rude to others					
38	I am someone who makes plans and follows through with them					
39	I am someone who gets nervous easily					
40	I am someone who likes to reflect, play with ideas					
41	I am someone who has few artistic interests					
42	I am someone who likes to cooperate with others					
43	I am easily distracted					
44	I am sophisticated in art, music, or literature					

Page **3** of **3**



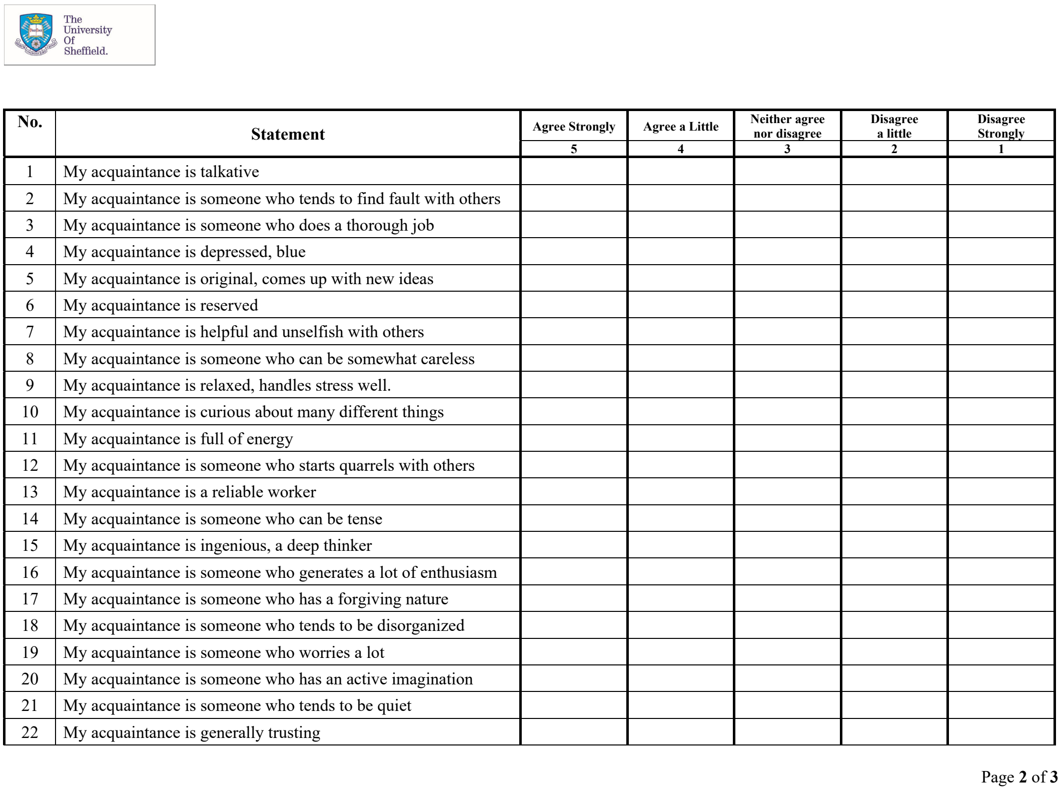
The University Of Sheffield.

Participant (Acquaintance) Personality Questionnaire
Automatic Personality Recognition

Code: __________ (***Filled by main researcher***)

Age:	
Gender:	□ **Male** □ **Female**
Ethnicity:	□ **Arab** □ **White** □ **Hispanic, Latino, Spanish** □ **Asian** □ **Black, African American** □ **Other**________
Nationality:	
Native Language:	
How long have you known the target?	
What is your relationship to the target?	
How well do you know the target?	**Not Very well** … **Very well** **1 2 3 4 5 6 7 8 9**

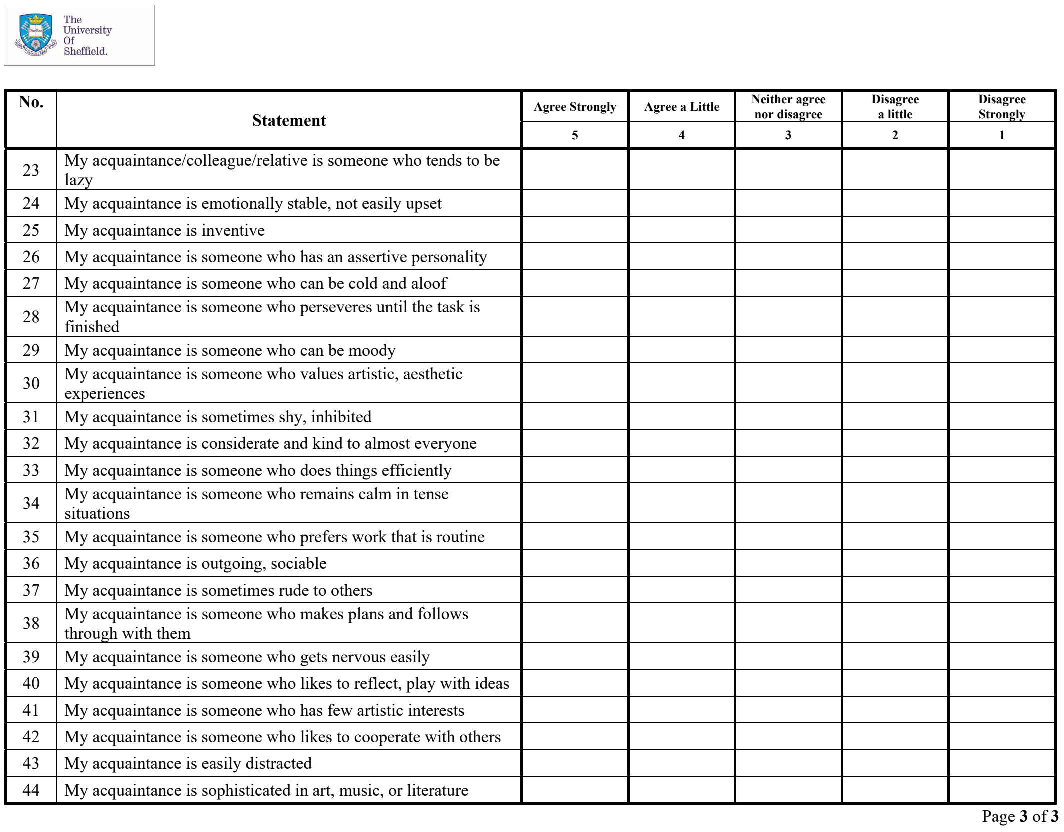
Page **1** of **3**



The
University
Of
Sheffield.

No.	Statement	Agree Strongly	Agree a Little	Neither agree nor disagree	Disagree a little	Disagree Strongly
		5	4	3	2	1
1	My acquaintance is talkative					
2	My acquaintance is someone who tends to find fault with others					
3	My acquaintance is someone who does a thorough job					
4	My acquaintance is depressed, blue					
5	My acquaintance is original, comes up with new ideas					
6	My acquaintance is reserved					
7	My acquaintance is helpful and unselfish with others					
8	My acquaintance is someone who can be somewhat careless					
9	My acquaintance is relaxed, handles stress well.					
10	My acquaintance is curious about many different things					
11	My acquaintance is full of energy					
12	My acquaintance is someone who starts quarrels with others					
13	My acquaintance is a reliable worker					
14	My acquaintance is someone who can be tense					
15	My acquaintance is ingenious, a deep thinker					
16	My acquaintance is someone who generates a lot of enthusiasm					
17	My acquaintance is someone who has a forgiving nature					
18	My acquaintance is someone who tends to be disorganized					
19	My acquaintance is someone who worries a lot					
20	My acquaintance is someone who has an active imagination					
21	My acquaintance is someone who tends to be quiet					
22	My acquaintance is generally trusting					

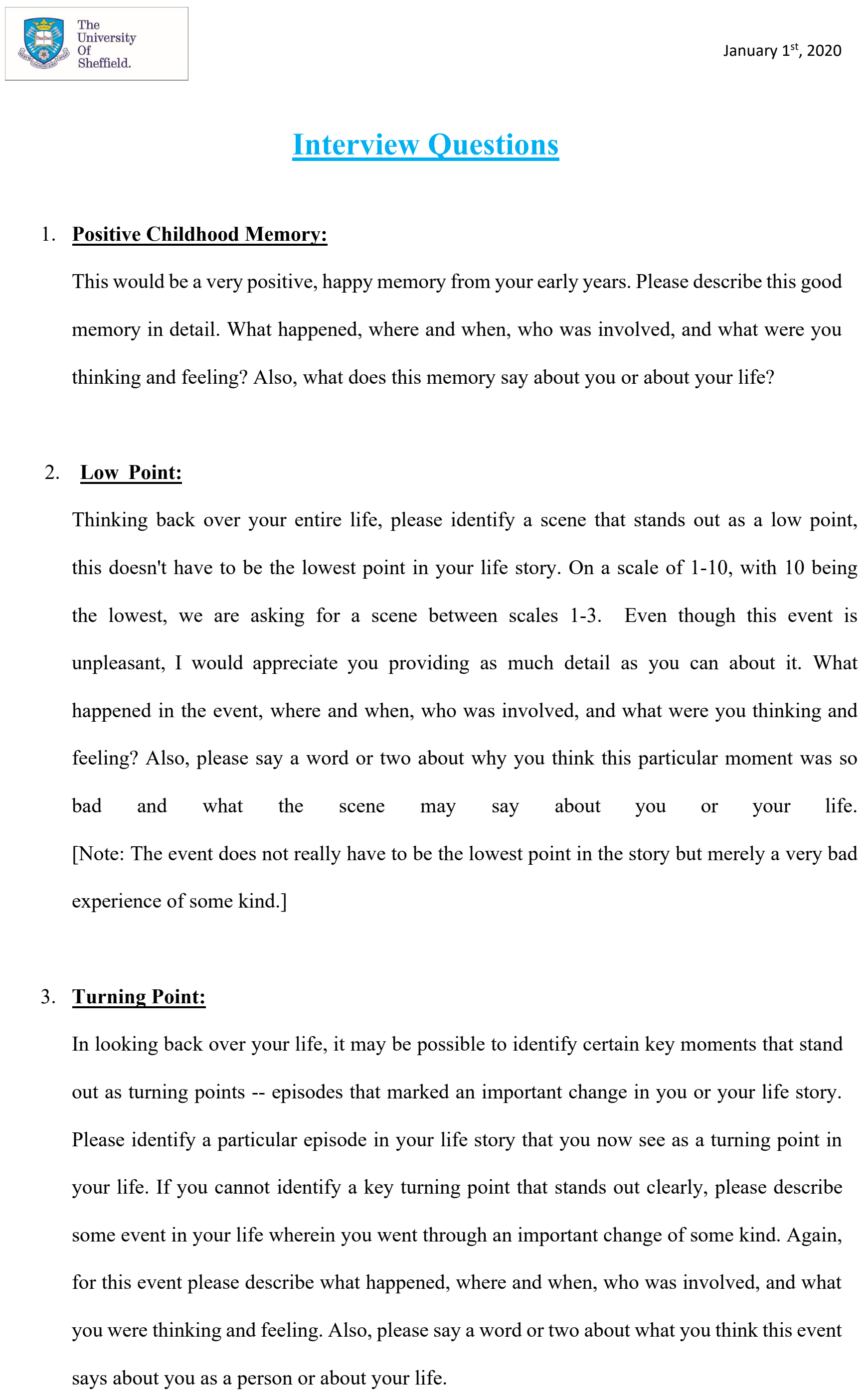
Page **2** of **3**



The University Of Sheffield.

No.	Statement	Agree Strongly	Agree a Little	Neither agree nor disagree	Disagree a little	Disagree Strongly
		5	4	3	2	1
23	My acquaintance/colleague/relative is someone who tends to be lazy					
24	My acquaintance is emotionally stable, not easily upset					
25	My acquaintance is inventive					
26	My acquaintance is someone who has an assertive personality					
27	My acquaintance is someone who can be cold and aloof					
28	My acquaintance is someone who perseveres until the task is finished					
29	My acquaintance is someone who can be moody					
30	My acquaintance is someone who values artistic, aesthetic experiences					
31	My acquaintance is sometimes shy, inhibited					
32	My acquaintance is considerate and kind to almost everyone					
33	My acquaintance is someone who does things efficiently					
34	My acquaintance is someone who remains calm in tense situations					
35	My acquaintance is someone who prefers work that is routine					
36	My acquaintance is outgoing, sociable					
37	My acquaintance is sometimes rude to others					
38	My acquaintance is someone who makes plans and follows through with them					
39	My acquaintance is someone who gets nervous easily					
40	My acquaintance is someone who likes to reflect, play with ideas					
41	My acquaintance is someone who has few artistic interests					
42	My acquaintance is someone who likes to cooperate with others					
43	My acquaintance is easily distracted					
44	My acquaintance is sophisticated in art, music, or literature					

Page **3** of **3**



The
University
Of
Sheffield.

January 1st, 2020

Interview Questions

1. **Positive Childhood Memory:**

   This would be a very positive, happy memory from your early years. Please describe this good memory in detail. What happened, where and when, who was involved, and what were you thinking and feeling? Also, what does this memory say about you or about your life?

2. **Low Point:**

   Thinking back over your entire life, please identify a scene that stands out as a low point, this doesn't have to be the lowest point in your life story. On a scale of 1-10, with 10 being the lowest, we are asking for a scene between scales 1-3. Even though this event is unpleasant, I would appreciate you providing as much detail as you can about it. What happened in the event, where and when, who was involved, and what were you thinking and feeling? Also, please say a word or two about why you think this particular moment was so bad and what the scene may say about you or your life. [Note: The event does not really have to be the lowest point in the story but merely a very bad experience of some kind.]

3. **Turning Point:**

   In looking back over your life, it may be possible to identify certain key moments that stand out as turning points -- episodes that marked an important change in you or your life story. Please identify a particular episode in your life story that you now see as a turning point in your life. If you cannot identify a key turning point that stands out clearly, please describe some event in your life wherein you went through an important change of some kind. Again, for this event please describe what happened, where and when, who was involved, and what you were thinking and feeling. Also, please say a word or two about what you think this event says about you as a person or about your life.

Appendix C: PTC Exploratory Data Figures and Tables

This appendix lists the exploratory data analysis tables and graphs which were derived from the new Personality Traits Corpus 2020.

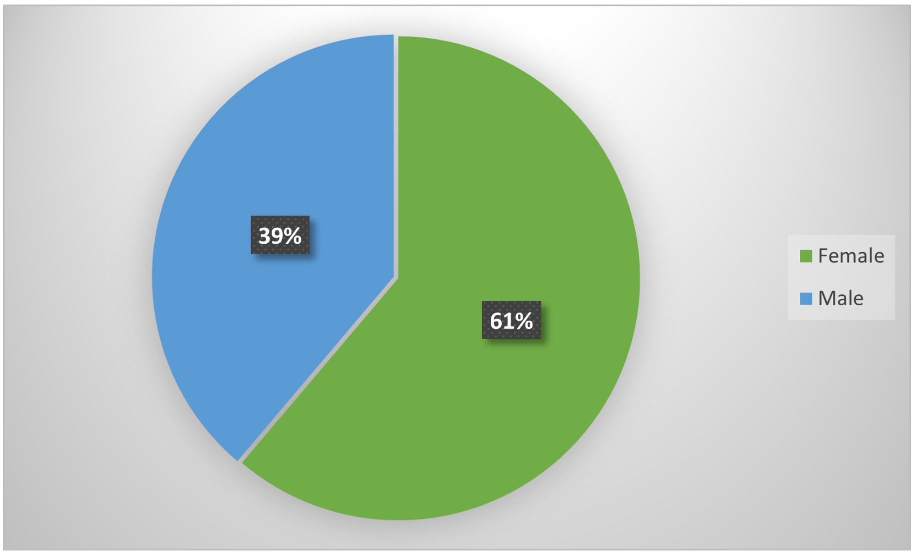


Figure C.1: *Target participant gender distribution.*

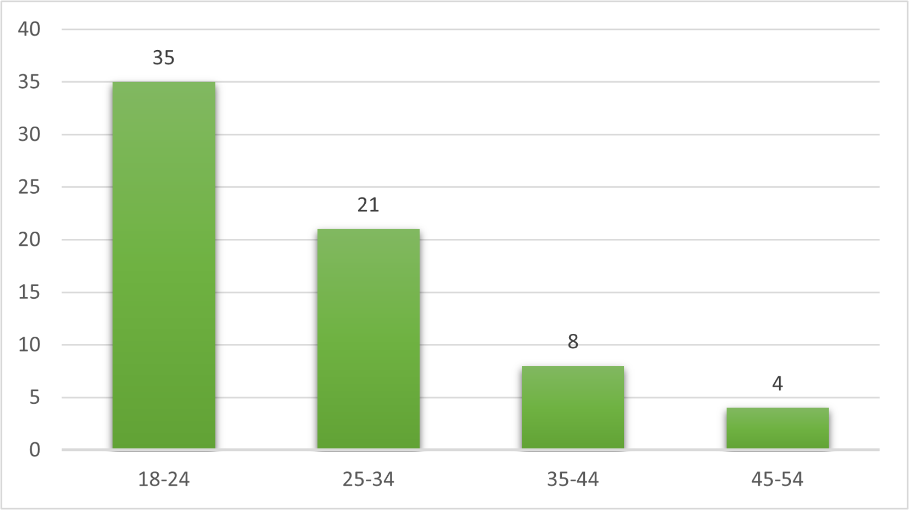


Figure C.2: *Target participant age group distribution.*

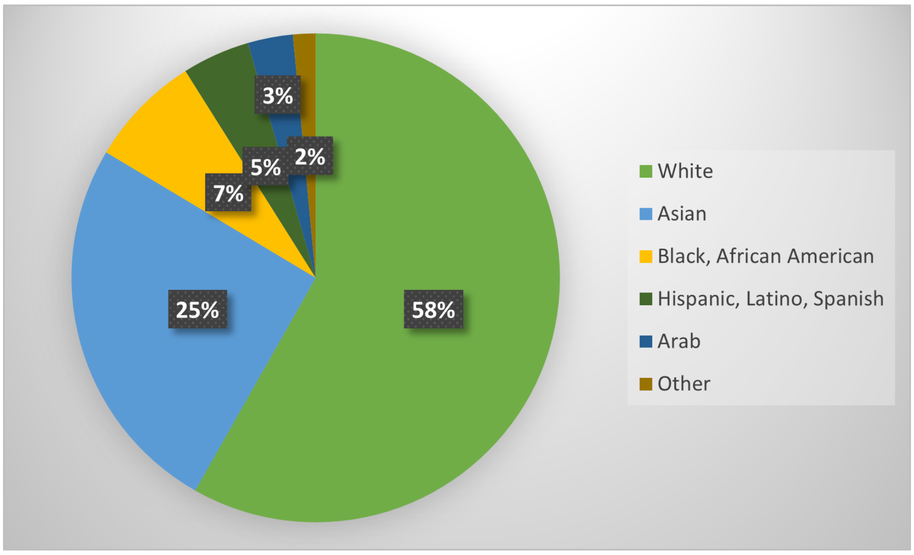


Figure C.3: *Target participant ethnicity distribution.*

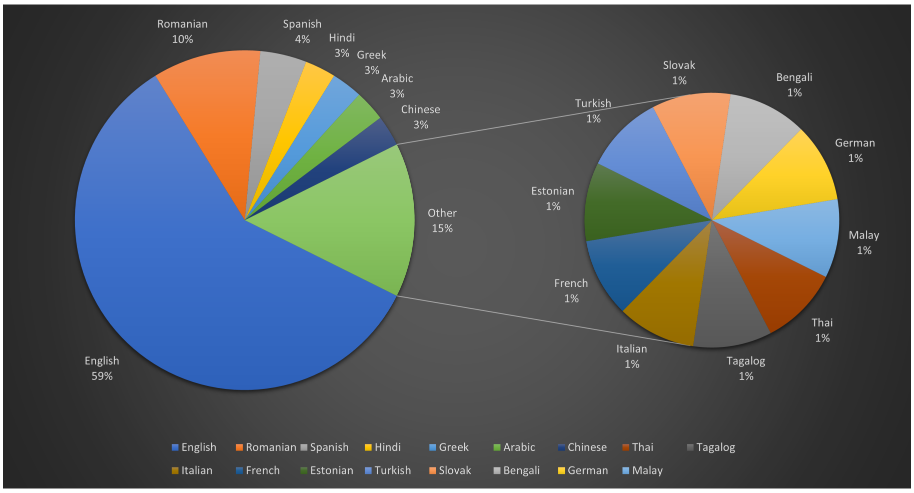


Figure C.4: *Target participant language distribution.*

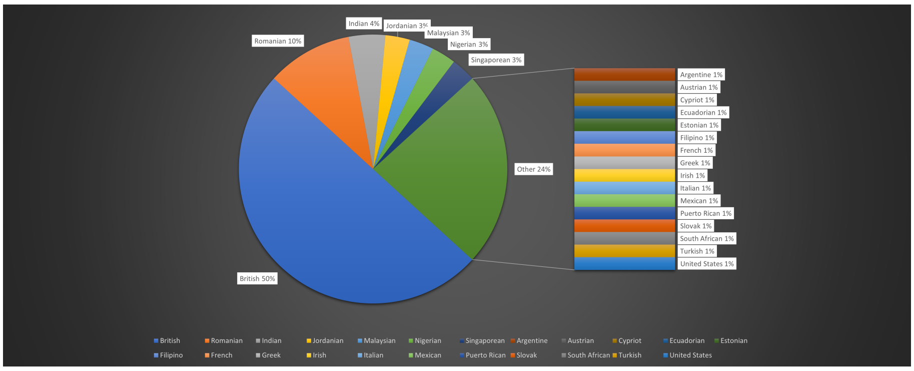


Figure C.5: *Target participant nationality distribution.*

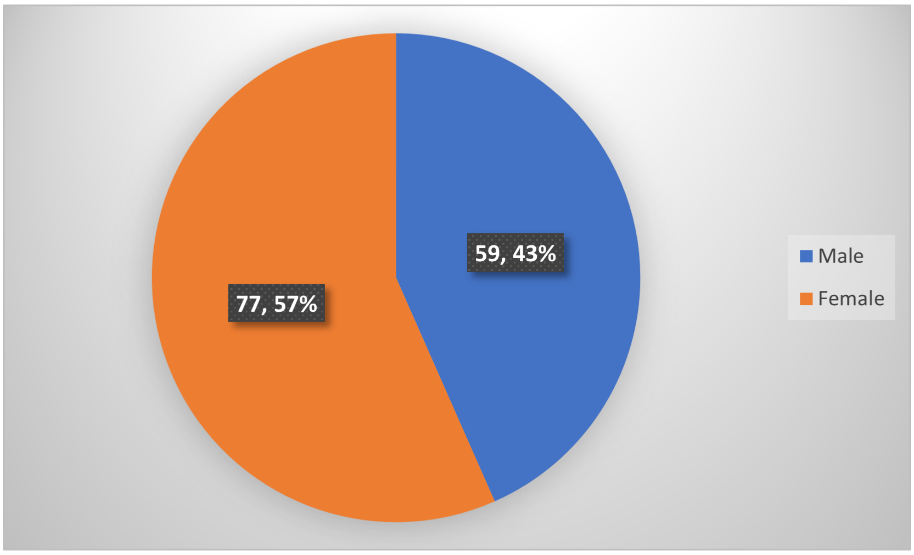


Figure C.6: *Acquaintance participant gender distribution.*

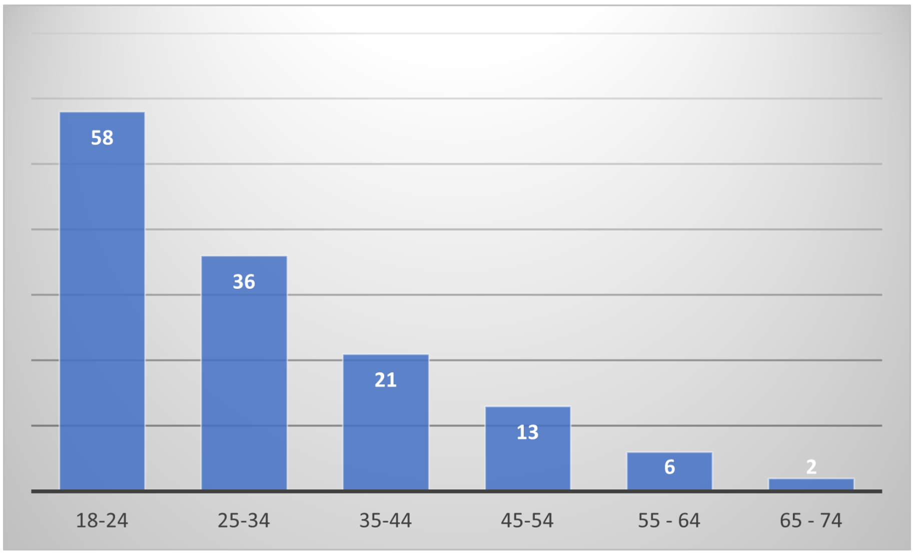


Figure C.7: *Acquaintance participant age group distribution.*

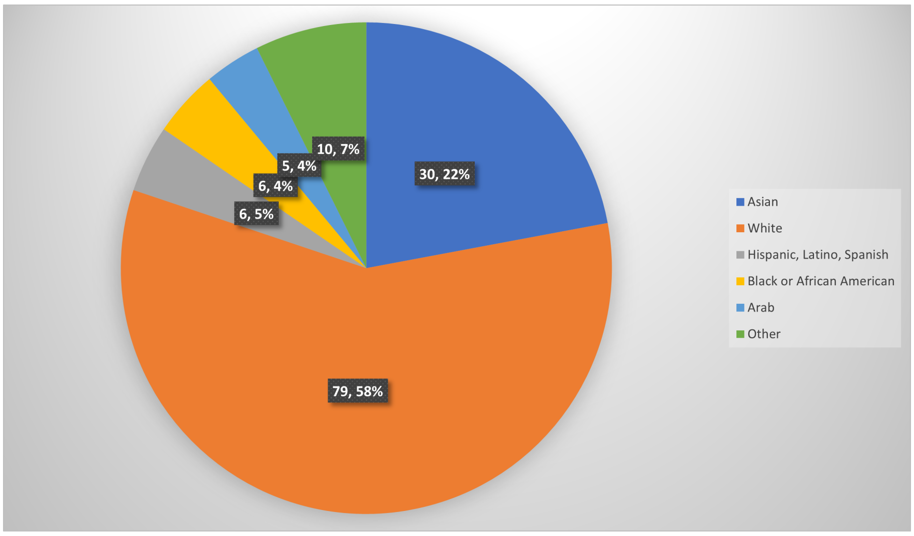


Figure C.8: *Acquaintance participant ethnicity distribution.*

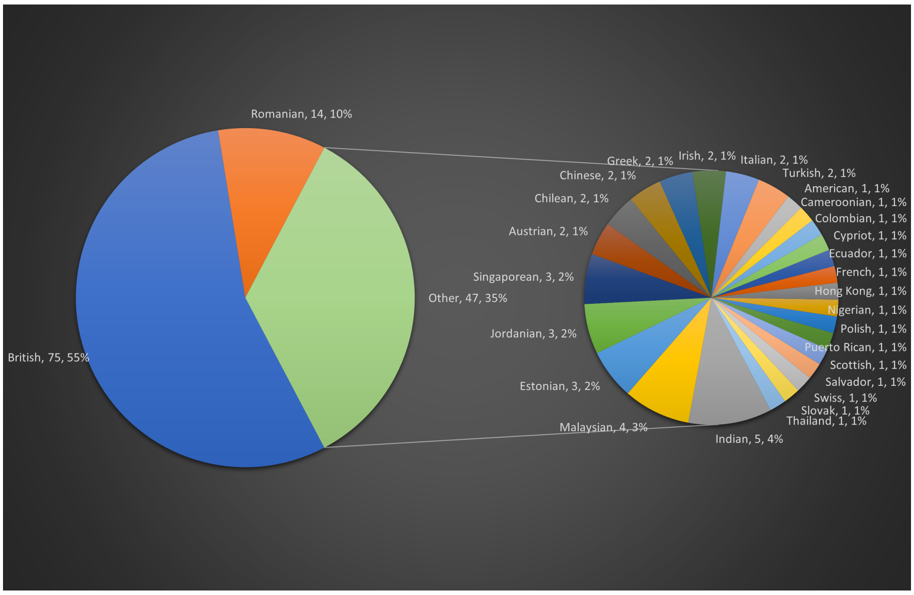


Figure C.9: *Acquaintance participant nationality distribution.*

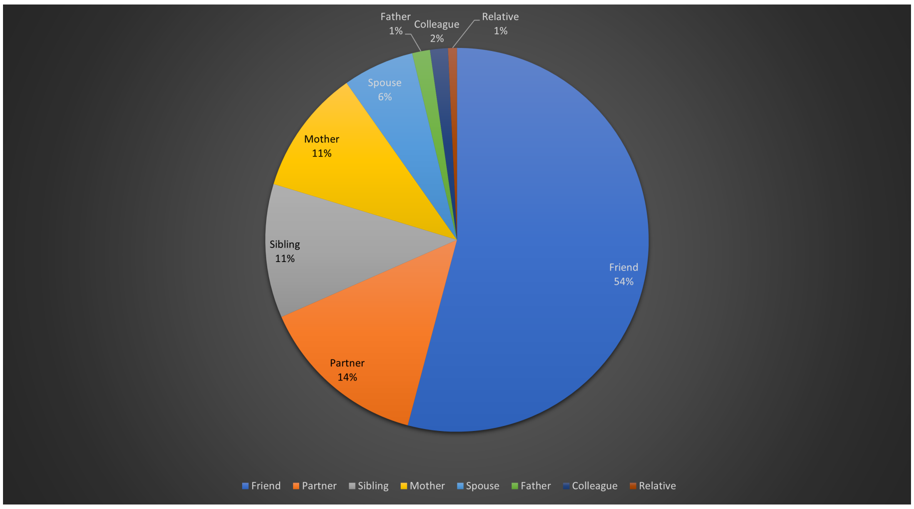


Figure C.10: *Acquaintance participant relation to target distribution.*

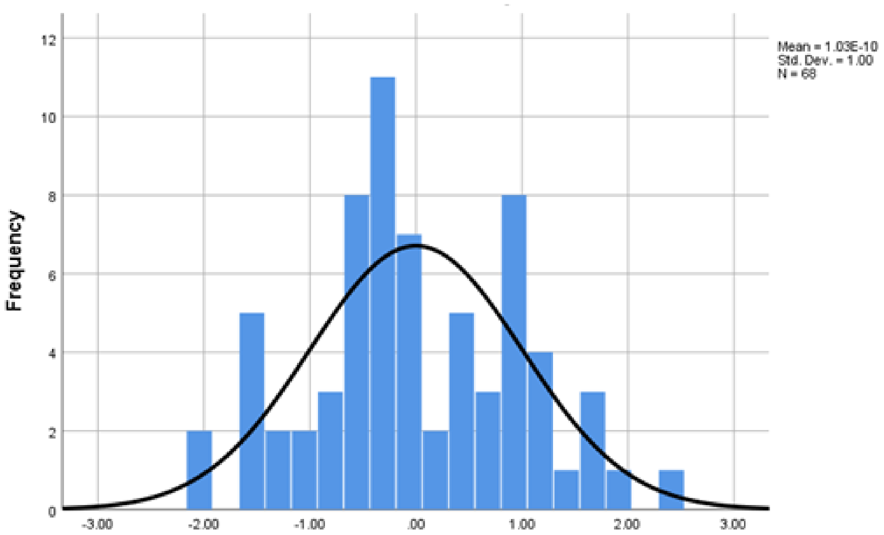


Figure C.11: *Openness histogram.*

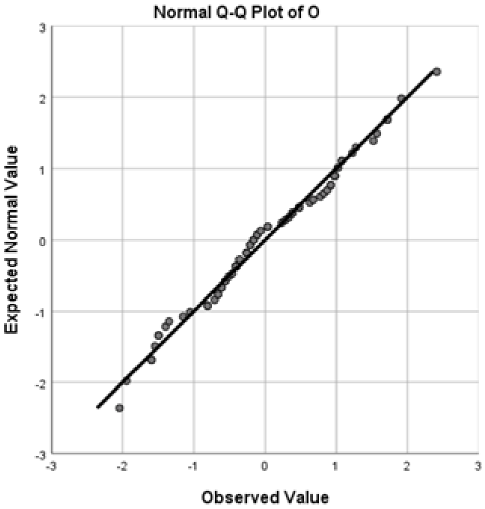


Figure C.12: *Openness QQ-plot.*

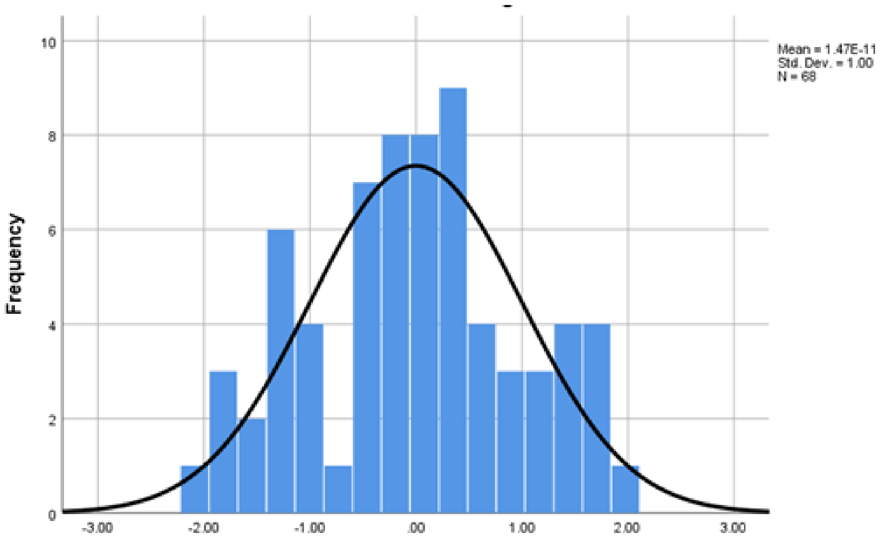


Figure C.13: *Conscientiousness histogram.*

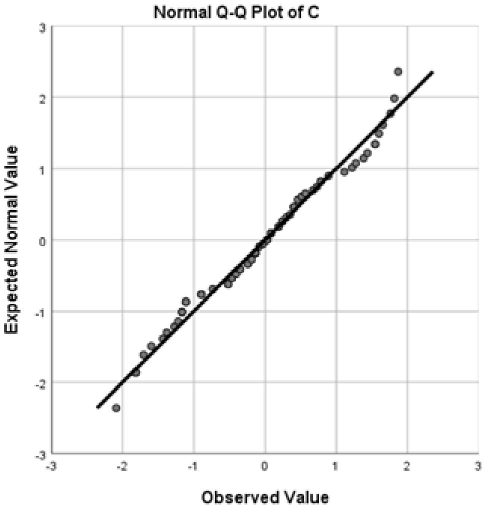


Figure C.14: *Conscientiousness QQ-plot.*

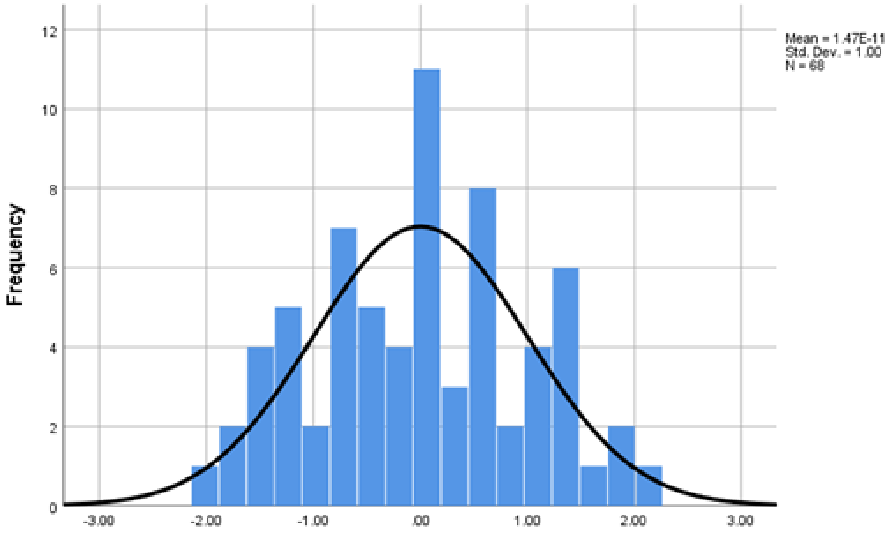


Figure C.15: *Extraversion histogram.*

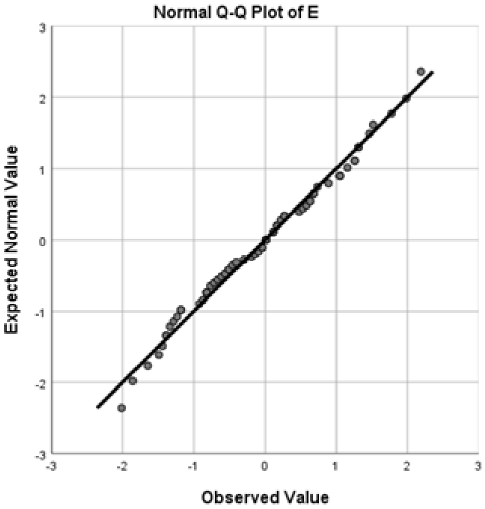


Figure C.16: *Extraversion QQ-plot.*

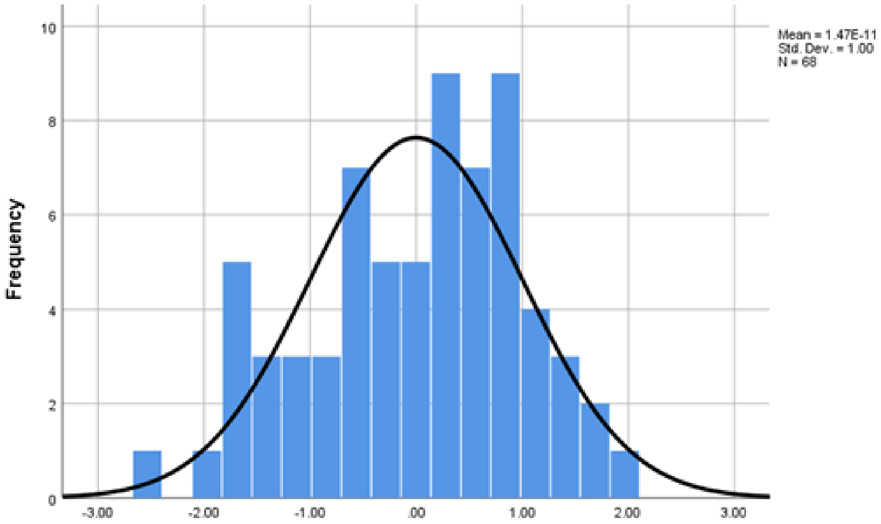


Figure C.17: *Agreeableness histogram.*

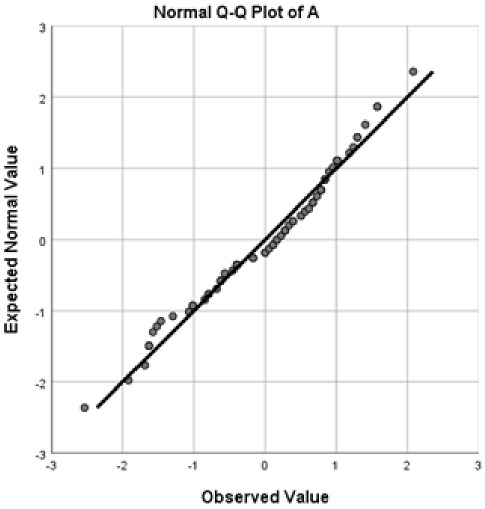


Figure C.18: *Agreeableness QQ-plot.*

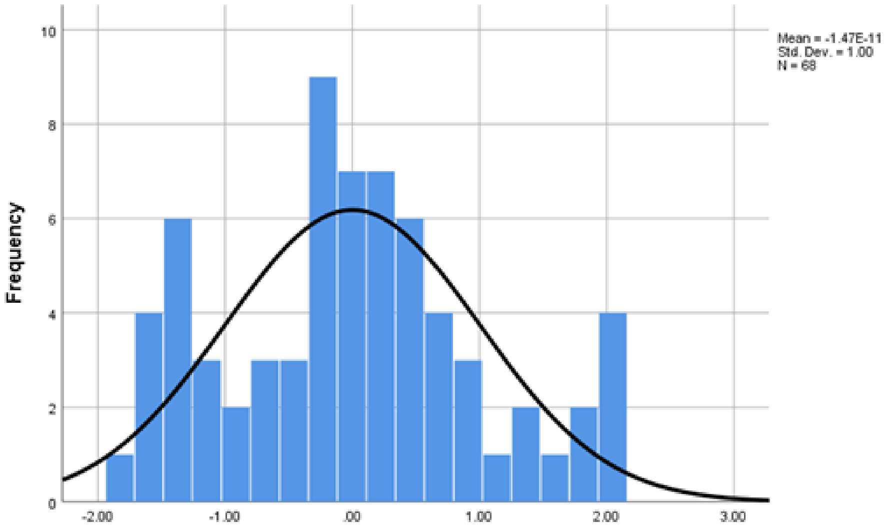


Figure C.19: *Neuroticism histogram.*

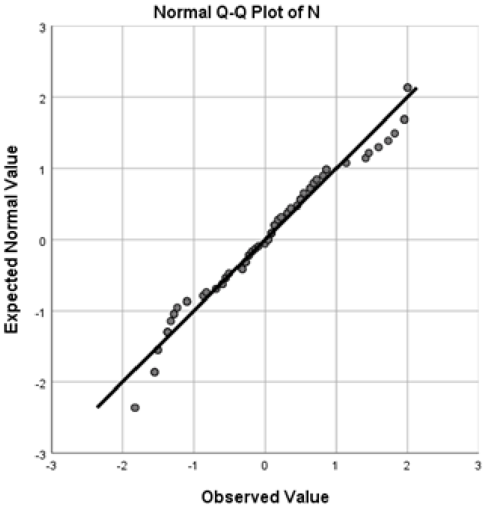


Figure C.20: *Neuroticism QQ-plot.*

Appendix D: PTC Experiments Figures and Tables

This appendix lists the tables and graphs which were derived from the experiments on the new Personality Traits Corpus.

Table D.1: *Openness trait prediction from gradually augmented dataset.*

	No Augment					10%					20%					30%					40%					50%					60%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	30.95%	48.12%	-0.13394993	0.696404549	0.411764706	30.95%	48.39%	-0.057894737	0.696404549	0.411764706	34.29%	50.75%	0.031578947	0.565017799	0.392156863	29.52%	46.98%	-0.10019152	0.807315441	0.431372549	32.86%	49.48%	-0.01950092	0.696404549	0.411764706	34.29%	50.75%	0.031578947	0.565	0.3922	34.29%	50.75%	0.0316	0.565	0.3922
Perceptron	43.33%	57.36%	0.145698559	0.427061712	0.37254902	39.52%	54.48%	0.113362577	0.427061712	0.37254902	46.19%	59.48%	0.22373999	0.191271766	0.333333333	49.52%	61.85%	0.277311536	0.111942958	0.31372549	50.48%	62.46%	0.272480484	0.111942958	0.31372549	42.86%	56.85%	0.171377655	0.2983	0.3529	50.95%	63.12%	0.3637	0.0595	0.2941
Neural Net	33.33%	50.14%	0.017859991	0.565017799	0.392156863	33.33%	50.28%	0.025649459	0.565017799	0.392156863	43.33%	57.50%	0.3	0.191271766	0.333333333	43.33%	57.50%	0.3	0.191271766	0.333333333	40.00%	55.00%	0.24178821	0.298302697	0.352941176	40.00%	55.00%	0.24178821	0.2983	0.3529	40.00%	55.00%	0.2418	0.2983	0.3529
Deep Learning	33.33%	50.00%	0	0.807315441	0.431372549	43.33%	57.50%	0.3	0.191271766	0.333333333	36.67%	52.50%	0.169669911	0.427061712	0.37254902	40.00%	55.00%	0.24178821	0.298302697	0.352941176	36.67%	52.50%	0.169669911	0.427061712	0.37254902	46.67%	60.00%	0.350115188	0.1119	0.3137	44.29%	58.12%	0.2547	0.1913	0.3333
SVM	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565	0.3922	33.33%	50.00%	None	0.565	0.3922
Naïve Bayes	33.33%	50.14%	0.017859991	0.565017799	0.392156863	28.10%	45.85%	-0.108597294	0.889981945	0.450980392	29.05%	46.33%	-0.09787413	0.889981945	0.450980392	29.05%	46.33%	-0.09787413	0.889981945	0.450980392	29.05%	46.33%	-0.09787413	0.889981945	0.450980392	29.05%	46.33%	-0.09787413	0.89	0.451	29.05%	46.33%	-0.098	0.89	0.451
Logistic Regression	36.67%	52.50%	0.169669911	0.427061712	0.37254902	46.67%	32.56%	0.601388889	0.111942958	0.31372549	40.00%	55.14%	0.210526316	0.298302697	0.352941176	40.00%	55.00%	0.24178821	0.298302697	0.352941176	40.00%	55.00%	0.24178821	0.298302697	0.352941176	40.00%	55.00%	0.24178821	0.2983	0.3529	40.00%	55.00%	0.2418	0.2983	0.3529
KNN	33.33%	50.00%	None	0.565017799	0.392156863	43.33%	57.50%	0.3	0.191271766	0.333333333	50.00%	62.50%	0.396591437	0.059497802	0.294117647	50.00%	62.50%	0.396591437	0.059497802	0.294117647	63.33%	72.50%	0.560429864	0.001645224	0.215686275	66.67%	75.00%	0.598740817	0.0005	0.1961	70.00%	77.50%	0.6373	0.0001	0.1765
Bagging	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565	0.3922	33.33%	50.00%	None	0.565	0.3922
Random Forest	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565	0.3922	33.33%	50.00%	None	0.565	0.3922
Ada Boost	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	34.29%	50.75%	0.031578947	0.565017799	0.392156863	34.29%	50.62%	0.025649459	0.565017799	0.392156863	30.95%	48.12%	-0.13394993	0.696404549	0.411764706	33.33%	50.00%	None	0.565	0.3922	33.33%	50.00%	None	0.565	0.3922
Linear SVC	36.67%	52.50%	0.050808678	0.696404549	0.411764706	44.29%	58.25%	0.243608489	0.191271766	0.333333333	45.24%	58.87%	0.229646058	0.191271766	0.333333333	47.62%	60.62%	0.309344112	0.111942958	0.31372549	37.14%	52.46%	0.054963674	0.565017799	0.392156863	47.62%	60.62%	0.309344112	0.1119	0.3137	44.29%	58.12%	0.2547	0.1913	0.3333
Passive Aggressive	28.57%	46.65%	-0.062746438	0.944108823	0.470588235	47.62%	60.89%	0.290520075	0.111942958	0.31372549	39.52%	54.48%	0.113109253	0.427061712	0.37254902	50.00%	62.50%	0.396591437	0.059497802	0.294117647	44.29%	58.12%	0.254653449	0.191271766	0.333333333	43.33%	57.50%	0.3	0.1913	0.3333	43.33%	57.50%	0.3	0.1913	0.3333
Ridge	33.33%	50.00%	None	0.565017799	0.392156863	36.67%	52.50%	0.169669911	0.427061712	0.37254902	36.67%	52.50%	0.169669911	0.427061712	0.37254902	36.67%	52.50%	0.169669911	0.427061712	0.37254902	36.67%	52.50%	0.169669911	0.427061712	0.37254902	36.67%	52.50%	0.169669911	0.4271	0.3725	36.67%	52.50%	0.1697	0.4271	0.3725
Gradient Boosting	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	30.95%	48.12%	-0.13394993	0.696404549	0.411764706	34.29%	50.62%	0.025649459	0.565017799	0.392156863	36.67%	52.64%	0.1336198	0.427061712	0.37254902	33.33%	50.00%	None	0.565	0.3922	33.33%	50.14%	0.0179	0.565	0.3922
LDA	26.67%	45.28%	-0.180861298	0.990333346	0.509803922	36.67%	52.50%	0.169669911	0.427061712	0.37254902	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565017799	0.392156863	33.33%	50.00%	None	0.565	0.3922	33.33%	50.00%	None	0.565	0.3922
QDA						41.90%	56.23%	0.127414535	0.565017799	0.392156863	59.05%	69.25%	0.382204191	0.028582762	0.274509804	51.90%	63.45%	0.262257407	0.191271766	0.333333333	57.62%	67.98%	0.357623736	0.059497802	0.294117647	55.24%	65.95%	0.323099771	0.1119	0.3137	55.71%	66.88%	0.3445	0.0595	0.2941
SGD	34.29%	50.62%	0.010624629	0.807315441	0.431372549	45.24%	58.73%	0.235192533	0.191271766	0.333333333	43.81%	57.60%	0.170766631	0.298302697	0.352941176	45.24%	58.73%	0.23637145	0.191271766	0.333333333	50.48%	62.46%	0.272935756	0.111942958	0.31372549	44.29%	58.12%	0.256356836	0.1913	0.3333	40.00%	55.00%	0.2418	0.2983	0.3529

Table D.2: *Conscientiousness trait prediction from gradually augmented dataset.*

	No Augment					10%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	39.32%	54.59%	0.125144928	0.297556089	0.388888889	28.67%	46.51%	-0.099013136	0.805907508	0.46296
Perceptron	24.94%	43.72%	-0.13273624	0.944420126	0.5	48.87%	61.60%	0.259541766	0.061213081	0.33333
Neural Net	33.80%	50.35%	0.01452329	0.562473141	0.425925926	36.32%	52.28%	0.09585851	0.425108744	0.40741
Deep Learning	36.11%	52.20%	0.09585851	0.425108744	0.407407407	39.36%	54.55%	0.149949673	0.297556089	0.38889
SVM	33.33%	50.00%	None	0.562473141	0.425925926	30.77%	48.05%	-0.138412803	0.694114022	0.44444
Naïve Bayes	33.33%	50.00%	None	0.562473141	0.425925926	36.11%	52.11%	0.150625698	0.425108744	0.40741
Logistic Regression	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.42593
KNN	28.67%	46.51%	-0.107707605	0.805907508	0.462962963	34.73%	50.93%	0.023665266	0.562473141	0.42593
Bagging	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.93%	0.023665266	0.562473141	0.42593
Random Forest	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.93%	0.023665266	0.562473141	0.42593
Ada Boost	36.11%	52.17%	0.113076259	0.425108744	0.407407407	36.36%	52.24%	0.154696662	0.425108744	0.40741
Linear SVC	27.97%	45.93%	-0.08877747	0.889587248	0.481481481	48.41%	61.34%	0.270294694	0.061213081	0.33333
Passive Aggressive	30.03%	47.54%	-0.050831517	0.805907508	0.462962963	45.84%	59.39%	0.216757373	0.113374316	0.35185
Ridge	33.33%	50.00%	None	0.562473141	0.425925926	36.36%	52.24%	0.154696662	0.425108744	0.40741
Gradient Boosting	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.93%	0.023665266	0.562473141	0.42593
LDA	33.33%	50.03%	0.004070965	0.562473141	0.425925926	33.33%	50.93%	0.023665266	0.562473141	0.42593
QDA						24.61%	43.59%	-0.125144928	0.944420126	0.5
SGD	27.97%	45.93%	-0.08877747	0.889587248	0.481481481	49.13%	61.81%	0.25382854	0.061213081	0.33333

	20%					30%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	33.55%	50.17%	0.00726	0.56247	0.42593	36.58%	52.38%	0.09586	0.42511	0.40741
Perceptron	43.49%	57.50%	0.16143	0.19182	0.37037	50.76%	63.12%	0.35364	0.03009	0.31481
Neural Net	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42511	0.40741
Deep Learning	38.89%	54.26%	0.18154	0.29756	0.38889	38.89%	54.23%	0.21613	0.29756	0.38889
SVM	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42511	0.40741
Naïve Bayes	38.89%	54.23%	0.21613	0.29756	0.38889	38.89%	54.23%	0.21613	0.29756	0.38889
Logistic Regression	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42511	0.40741
KNN	37.76%	53.17%	0.08368	0.42511	0.40741	40.79%	55.41%	0.13905	0.29756	0.38889
Bagging	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
Random Forest	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
Ada Boost	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
Linear SVC	56.06%	67.22%	0.43491	0.00542	0.27778	50.76%	63.12%	0.35364	0.03009	0.31481
Passive Aggressive	50.51%	63.02%	0.34103	0.03009	0.31481	44.70%	58.64%	0.26738	0.11337	0.35185
Ridge	36.11%	52.11%	0.15063	0.42511	0.40741	38.89%	54.23%	0.21613	0.29756	0.38889
Gradient Boosting	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
LDA	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
QDA	33.20%	49.92%	-0.0012	0.69411	0.44444	39.34%	54.57%	0.09586	0.42511	0.40741
SGD	48.15%	61.18%	0.26623	0.06121	0.33333	37.04%	52.70%	0.07426	0.42511	0.40741

	40%					50%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	33.55%	50.17%	0.00726	0.56247	0.42593	40.07%	54.91%	0.13469	0.29755609	0.38888889
Perceptron	36.32%	52.28%	0.08515	0.42511	0.40741	36.11%	52.14%	0.11098	0.42510874	0.40740741
Neural Net	36.11%	52.11%	0.15063	0.42511	0.40741	38.89%	54.23%	0.21613	0.29755609	0.38888889
Deep Learning	38.89%	54.23%	0.21613	0.29756	0.38889	38.89%	54.23%	0.21613	0.29755609	0.38888889
SVM	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42510874	0.40740741
Naïve Bayes	38.89%	54.23%	0.21613	0.29756	0.38889	38.89%	54.23%	0.21613	0.29755609	0.38888889
Logistic Regression	38.89%	54.23%	0.21613	0.29756	0.38889	36.11%	52.11%	0.15063	0.42510874	0.40740741
KNN	43.82%	57.65%	0.19119	0.19182	0.37037	46.60%	59.71%	0.23936	0.11337432	0.35185185
Bagging	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247314	0.42592593
Random Forest	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247314	0.42592593
Ada Boost	36.11%	52.11%	0.15063	0.42511	0.40741	30.77%	48.08%	-0.0957	0.69411402	0.44444444
Linear SVC	41.92%	56.50%	0.23778	0.19182	0.37037	38.89%	54.23%	0.21613	0.29755609	0.38888889
Passive Aggressive	38.89%	54.26%	0.18154	0.29756	0.38889	38.89%	54.23%	0.21613	0.29755609	0.38888889
Ridge	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42510874	0.40740741
Gradient Boosting	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.03%	0.00407	0.56247314	0.42592593
LDA	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247314	0.42592593
QDA	44.21%	58.00%	0.16156	0.19182	0.37037	45.82%	59.18%	0.18642	0.19181679	0.37037037
SGD	44.95%	58.74%	0.28661	0.11337	0.35185	36.11%	52.11%	0.15063	0.42510874	0.40740741

	60%					70%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	37.04%	52.70%	0.07426	0.42511	0.40741	37.30%	53.08%	0.08076	0.42511	0.40741
Perceptron	38.89%	54.23%	0.21613	0.29756	0.38889	36.58%	52.40%	0.09825	0.42511	0.40741
Neural Net	33.33%	50.06%	0.00814	0.56247	0.42593	36.11%	52.11%	0.15063	0.42511	0.40741
Deep Learning	38.89%	54.23%	0.21613	0.29756	0.38889	38.89%	54.23%	0.21613	0.29756	0.38889
SVM	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42511	0.40741
Naïve Bayes	36.11%	52.11%	0.15063	0.42511	0.40741	33.55%	50.14%	0.00584	0.56247	0.42593
Logistic Regression	36.11%	52.11%	0.15063	0.42511	0.40741	36.11%	52.11%	0.15063	0.42511	0.40741
KNN	55.19%	66.17%	0.36515	0.01343	0.2963	55.19%	66.17%	0.36515	0.01343	0.2963
Bagging	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
Random Forest	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
Ada Boost	36.58%	52.40%	0.09825	0.42511	0.40741	33.33%	50.00%	None	0.56247	0.42593
Linear SVC	36.79%	52.54%	0.08162	0.42511	0.40741	39.36%	54.52%	0.16351	0.29756	0.38889
Passive Aggressive	39.36%	54.49%	0.16203	0.29756	0.38889	39.14%	54.35%	0.21745	0.29756	0.38889
Ridge	41.92%	56.47%	0.2684	0.19182	0.37037	41.92%	56.47%	0.2684	0.19182	0.37037
Gradient Boosting	30.77%	48.05%	-0.1384	0.69411	0.44444	33.33%	50.03%	0.00407	0.56247	0.42593
LDA	33.33%	50.00%	None	0.56247	0.42593	33.33%	50.00%	None	0.56247	0.42593
QDA	55.32%	66.39%	0.33301	0.01343	0.2963	47.49%	60.62%	0.2174	0.11337	0.35185
SGD	39.36%	54.52%	0.16351	0.29756	0.38889	38.89%	54.23%	0.21613	0.29756	0.38889

Table D.3: *Extraversion trait prediction from gradually augmented dataset.*

	No Augment					10%					20%					30%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	28.21%	46.11%	-0.144755574	0.805907508	0.462962963	35.41%	51.30%	0.024904771	0.562473141	0.4259	36.83%	52.53%	0.1017	0.4251	0.4074	36.83%	52.50%	0.0983	0.4251	0.4074
Perceptron	41.61%	56.41%	0.135430134	0.297556089	0.388888889	37.47%	53.09%	0.069124189	0.425108744	0.4074	40.25%	55.03%	0.1218	0.2976	0.3889	43.32%	57.28%	0.1812	0.1918	0.3704
Neural Net	25.64%	44.15%	-0.246895937	0.889587248	0.481481481	30.77%	48.05%	-0.138412803	0.694114022	0.4444	36.36%	52.24%	0.1547	0.4251	0.4074	33.33%	50.00%	None	0.5625	0.4259
Deep Learning	34.48%	50.90%	0.022866104	0.562473141	0.425925926	39.61%	54.64%	0.166235132	0.297556089	0.3889	36.83%	52.53%	0.1017	0.4251	0.4074	33.80%	50.29%	0.0117	0.5625	0.4259
SVM	33.33%	50.00%	None	0.562473141	0.425925926	30.77%	48.02%	-0.142483768	0.694114022	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444
Naïve Bayes	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Logistic Regression	33.33%	50.00%	None	0.562473141	0.425925926	30.77%	48.02%	-0.142483768	0.694114022	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444
KNN	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	36.11%	52.11%	0.1506	0.4251	0.4074	44.70%	58.58%	0.3135	0.1134	0.3519
Bagging	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Random Forest	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Ada Boost	28.21%	46.10%	-0.19860301	0.805907508	0.462962963	28.42%	46.18%	-0.144755574	0.805907508	0.463	33.33%	50.00%	None	0.5625	0.4259	36.36%	52.27%	0.1131	0.4251	0.4074
Linear SVC	41.61%	56.44%	0.136286922	0.297556089	0.388888889	34.69%	50.83%	0.01652561	0.562473141	0.4259	42.60%	56.90%	0.185	0.1918	0.3704	39.57%	54.63%	0.1399	0.2976	0.3889
Passive Aggressive	50.91%	63.15%	0.27080029	0.061213081	0.333333333	37.26%	52.84%	0.070609423	0.425108744	0.4074	39.82%	54.75%	0.1405	0.2976	0.3889	33.76%	50.27%	0.0085	0.5625	0.4259
Ridge	33.33%	50.00%	None	0.562473141	0.425925926	30.77%	48.02%	-0.142483768	0.694114022	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444	30.77%	48.02%	-0.142	0.6941	0.4444
Gradient Boosting	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	33.33%	50.12%	0.0096	0.5625	0.4259	33.33%	50.03%	0.0041	0.5625	0.4259
LDA	33.33%	50.00%	None	0.562473141	0.425925926	33.33%	50.00%	None	0.562473141	0.4259	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
QDA						36.23%	52.24%	0.04673012	0.562473141	0.4259	40.17%	55.43%	0.1231	0.2976	0.3889	37.70%	53.23%	0.0737	0.5625	0.4259
SGD	44.64%	58.56%	0.175311589	0.191816787	0.37037037	37.47%	53.03%	0.069206402	0.425108744	0.4074	40.71%	55.38%	0.1162	0.2976	0.3889	48.15%	61.09%	0.2869	0.0612	0.3333

	40%					50%					60%					70%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	36.83%	52.50%	0.0983	0.4251	0.4074	33.80%	50.29%	0.0096	0.562473	0.42593	33.80%	50.29%	0.0096	0.5625	0.4259	30.77%	48.08%	-0.08	0.6941	0.4444
Perceptron	46.35%	59.55%	0.2313	0.1134	0.3519	49.38%	61.82%	0.2814	0.061213	0.33333	49.13%	61.67%	0.274	0.0612	0.3333	46.13%	59.42%	0.252	0.1134	0.3519
Neural Net	33.33%	50.00%	None	0.5625	0.4259	36.36%	52.24%	0.1547	0.425109	0.40741	33.33%	50.00%	None	0.5625	0.4259	39.39%	54.48%	0.222	0.2976	0.3889
Deep Learning	31.24%	48.34%	-0.073	0.6941	0.4444	34.27%	50.58%	0.017	0.562473	0.42593	34.27%	50.58%	0.017	0.5625	0.4259	34.27%	50.58%	0.017	0.5625	0.4259
SVM	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.562473	0.42593	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Naïve Bayes	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.562473	0.42593	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Logistic Regression	30.77%	48.05%	-0.138	0.6941	0.4444	31.24%	48.34%	-0.073	0.694114	0.44444	33.80%	50.29%	0.0117	0.5625	0.4259	33.80%	50.29%	0.0117	0.5625	0.4259
KNN	50.51%	62.93%	0.3935	0.0301	0.3148	56.57%	67.41%	0.4664	0.005423	0.27778	59.34%	69.53%	0.5005	0.002	0.2593	59.34%	69.53%	0.5005	0.002	0.2593
Bagging	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.562473	0.42593	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Random Forest	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.562473	0.42593	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
Ada Boost	36.36%	52.24%	0.1547	0.4251	0.4074	36.36%	52.24%	0.1547	0.425109	0.40741	36.36%	52.24%	0.1547	0.4251	0.4074	33.33%	50.00%	None	0.5625	0.4259
Linear SVC	43.32%	57.31%	0.1834	0.1918	0.3704	40.54%	55.23%	0.1379	0.297556	0.38889	31.70%	48.63%	-0.05	0.6941	0.4444	31.70%	48.63%	-0.05	0.6941	0.4444
Passive Aggressive	37.51%	52.96%	0.0792	0.4251	0.4074	34.73%	50.87%	0.0215	0.562473	0.42593	34.73%	50.87%	0.0215	0.5625	0.4259	31.70%	48.63%	-0.05	0.6941	0.4444
Ridge	30.77%	48.05%	-0.138	0.6941	0.4444	28.21%	46.10%	-0.199	0.805908	0.46296	28.21%	46.10%	-0.199	0.8059	0.463	37.30%	52.82%	0.089	0.4251	0.4074
Gradient Boosting	36.83%	52.50%	0.0983	0.4251	0.4074	33.33%	50.03%	0.0041	0.562473	0.42593	34.01%	50.43%	0.0115	0.5625	0.4259	36.11%	52.14%	0.111	0.4251	0.4074
LDA	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.562473	0.42593	33.33%	50.00%	None	0.5625	0.4259	33.33%	50.00%	None	0.5625	0.4259
QDA	41.10%	56.04%	0.1269	0.2976	0.3889	40.50%	55.62%	0.1243	0.297556	0.38889	61.81%	71.23%	0.4252	0.002	0.2593	37.65%	53.33%	0.0726	0.4251	0.4074
SGD	37.04%	52.70%	0.0859	0.4251	0.4074	37.30%	52.82%	0.089	0.425109	0.40741	40.07%	54.94%	0.147	0.2976	0.3889	37.30%	52.82%	0.089	0.4251	0.4074

Table D.4: *Agreeableness trait prediction from gradually augmented dataset.*

	No Augment					10%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	33.73%	50.41%	0.017915797	0.563860351	0.4	46.30%	59.66%	0.3394476	0.114313	0.3238
Perceptron	33.47%	49.44%	-0.024968964	0.887745856	0.457142857	43.25%	56.99%	0.1578992	0.299739	0.3619
Neural Net	33.33%	50.00%	None	0.563860351	0.4	33.73%	50.35%	0.0213264	0.56386	0.4
Deep Learning	41.67%	56.70%	0.256692457	0.193483497	0.342857143	37.83%	53.50%	0.109062	0.427246	0.381
SVM	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
Naïve Bayes	33.33%	50.00%	None	0.563860351	0.4	33.73%	50.27%	0.0122785	0.56386	0.4
Logistic Regression	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
KNN	30.95%	48.32%	-0.07674057	0.694265018	0.419047619	36.51%	52.53%	0.11084	0.427246	0.381
Bagging	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
Random Forest	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
Ada Boost	33.33%	50.00%	None	0.563860351	0.4	37.04%	52.80%	0.138133	0.427246	0.381
Linear SVC	35.85%	51.36%	0.016129082	0.804825497	0.438095238	41.53%	55.99%	0.1737597	0.299739	0.3619
Passive Aggressive	30.95%	48.37%	-0.030000375	0.887745856	0.457142857	44.31%	58.18%	0.2265148	0.193483	0.3429
Ridge	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
Gradient Boosting	33.33%	50.00%	None	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
LDA	34.66%	50.88%	0.030229098	0.563860351	0.4	33.33%	50.00%	None	0.56386	0.4
QDA						31.48%	49.18%	-0.005	0.804825	0.4381
SGD	41.01%	55.37%	0.099750934	0.563860351	0.4	34.39%	50.33%	-0.003129	0.694265	0.419

	20%					30%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	35.05%	51.46%	0.045	0.5639	0.4	38.76%	54.12%	0.1176	0.4272	0.381
Perceptron	40.61%	55.45%	0.1627	0.2997	0.3619	34.52%	50.80%	0.0227	0.5639	0.4
Neural Net	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Deep Learning	40.21%	55.10%	0.1792	0.2997	0.3619	37.43%	53.00%	0.1184	0.4272	0.381
SVM	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Naïve Bayes	31.35%	48.35%	-0.074	0.6943	0.419	31.35%	48.35%	-0.074	0.6943	0.419
Logistic Regression	30.95%	48.08%	-0.141	0.6943	0.419	30.95%	48.17%	-0.128	0.6943	0.419
KNN	43.92%	57.82%	0.2476	0.1935	0.3429	54.10%	65.30%	0.3954	0.03	0.2857
Bagging	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Random Forest	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Ada Boost	42.06%	56.88%	0.2212	0.1935	0.3429	27.51%	45.78%	-0.11	0.8877	0.4571
Linear SVC	45.77%	59.54%	0.2921	0.1143	0.3238	46.69%	60.01%	0.2968	0.1143	0.3238
Passive Aggressive	42.06%	56.82%	0.2342	0.1935	0.3429	47.09%	60.27%	0.2707	0.1143	0.3238
Ridge	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Gradient Boosting	33.33%	50.00%	None	0.5639	0.4	30.95%	48.24%	-0.085	0.6943	0.419
LDA	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
QDA	39.95%	54.66%	0.0901	0.6943	0.419	47.35%	59.73%	0.1844	0.4272	0.381
SGD	40.61%	55.36%	0.1584	0.2997	0.3619	47.49%	60.54%	0.256	0.1143	0.3238

	40%					50%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	37.83%	53.64%	0.1031	0.4272	0.381	34.13%	51.06%	0.041	0.56386	0.4
Perceptron	46.16%	59.73%	0.2655	0.1143	0.3238	38.89%	54.37%	0.2272	0.299739	0.361905
Neural Net	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
Deep Learning	39.81%	54.83%	0.2346	0.2997	0.3619	33.33%	50.00%	None	0.56386	0.4
SVM	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
Naïve Bayes	30.95%	48.08%	-0.141	0.6943	0.419	30.95%	48.08%	-0.141	0.694265	0.419048
Logistic Regression	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
KNN	60.58%	70.13%	0.4791	0.0053	0.2476	64.29%	72.77%	0.5225	0.00187	0.228571
Bagging	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
Random Forest	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
Ada Boost	39.29%	54.63%	0.1747	0.2997	0.3619	33.33%	50.00%	None	0.56386	0.4
Linear SVC	50.93%	63.56%	0.4194	0.03	0.2857	38.89%	54.37%	0.2272	0.299739	0.361905
Passive Aggressive	44.84%	59.00%	0.2874	0.1143	0.3238	38.89%	54.37%	0.2272	0.299739	0.361905
Ridge	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
Gradient Boosting	42.59%	57.01%	0.2872	0.1935	0.3429	42.59%	57.01%	0.2872	0.193483	0.342857
LDA	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.56386	0.4
QDA	45.77%	59.15%	0.1799	0.2997	0.3619	38.10%	52.97%	0.0487	0.804825	0.438095
SGD	40.21%	55.10%	0.1792	0.2997	0.3619	41.67%	56.55%	0.2825	0.193483	0.342857

	60%					70%				
Classifier	UAR	AUC	MCC	P-Value	Error Rate	UAR	AUC	MCC	P-Value	Error Rate
Decision Tree	32.54%	49.15%	-0.028	0.6943	0.419	32.54%	49.15%	-0.028	0.6943	0.419
Perceptron	39.29%	54.72%	0.1792	0.2997	0.3619	42.06%	56.90%	0.2371	0.1935	0.3429
Neural Net	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Deep Learning	30.95%	48.17%	-0.128	0.6943	0.419	30.95%	48.17%	-0.128	0.6943	0.419
SVM	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Naïve Bayes	28.57%	46.17%	-0.203	0.8048	0.4381	28.57%	46.17%	-0.203	0.8048	0.4381
Logistic Regression	30.95%	48.08%	-0.141	0.6943	0.419	33.33%	50.00%	None	0.5639	0.4
KNN	67.06%	74.95%	0.5586	0.0006	0.2095	67.06%	74.95%	0.5586	0.0006	0.2095
Bagging	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Random Forest	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Ada Boost	36.11%	52.18%	0.1582	0.4272	0.381	33.33%	50.00%	None	0.5639	0.4
Linear SVC	36.51%	52.53%	0.1108	0.4272	0.381	30.95%	48.17%	-0.128	0.6943	0.419
Passive Aggressive	33.73%	50.35%	0.0213	0.5639	0.4	30.95%	48.17%	-0.128	0.6943	0.419
Ridge	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
Gradient Boosting	38.89%	54.52%	0.199	0.2997	0.3619	42.59%	57.01%	0.2872	0.1935	0.3429
LDA	33.33%	50.00%	None	0.5639	0.4	33.33%	50.00%	None	0.5639	0.4
QDA	45.37%	58.68%	0.1693	0.4272	0.381	43.78%	57.62%	0.151	0.4272	0.381
SGD	36.51%	52.53%	0.1108	0.4272	0.381	36.51%	52.53%	0.1108	0.4272	0.381

Table D.5: *Neuroticism trait prediction from gradually augmented dataset.*

	No Augment					10%					20%					30%					40%					50%					60%					70%				
Classifier	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**	**UAR**	**AUC**	**MCC**	**P-Value**	**Error Rate**
Decision Tree	30.77%	47.96%	-0.149501678	0.695897937	0.438095238	30.77%	47.96%	-0.0782843	0.695897937	0.4381	34.47%	50.80%	0.0265	0.5634	0.419	37.04%	52.73%	0.1082	0.425	0.4	37.04%	52.73%	0.1082	0.425	0.4	37.04%	52.61%	0.0981	0.424971	0.4	37.04%	52.61%	0.0981	0.425	0.4	37.04%	52.61%	0.0981	0.425	0.4
Perceptron	33.90%	49.70%	-0.022222493	0.807973499	0.457142857	43.02%	56.59%	0.126599585	0.296388919	0.381	51.85%	63.40%	0.2591	0.0596	0.3238	51.85%	63.28%	0.273	0.0596	0.3238	60.68%	69.85%	0.4489	0.005	0.2667	55.56%	65.77%	0.3601	0.028892	0.3047619	43.02%	56.48%	0.1675	0.2964	0.381	39.32%	53.87%	0.1024	0.425	0.4
Neural Net	39.60%	54.77%	0.196245201	0.296388919	0.380952381	37.04%	52.73%	0.131829914	0.424971477	0.4	37.04%	52.61%	0.1666	0.425	0.4	37.04%	52.61%	0.1666	0.425	0.4	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419
Deep Learning	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	37.04%	52.61%	0.1666	0.425	0.4	37.04%	52.61%	0.1666	0.425	0.4	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	30.77%	47.96%	-0.15	0.6959	0.4381	33.33%	50.00%	None	0.5634	0.419
SVM	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419
Naïve Bayes	33.33%	50.00%	None	0.563445001	0.419047619	30.77%	47.96%	-0.14950168	0.695897937	0.4381	33.33%	50.00%	0	0.5634	0.419	33.33%	50.00%	0	0.5634	0.419	35.90%	52.04%	0.1495	0.425	0.4	35.90%	52.04%	0.1495	0.424971	0.4	33.33%	50.00%	0	0.5634	0.419	35.90%	52.04%	0.1495	0.425	0.4
Logistic Regression	33.33%	50.00%	None	0.563445001	0.419047619	30.77%	48.08%	-0.13241577	0.695897937	0.4381	40.74%	55.34%	0.1812	0.2964	0.381	38.18%	53.30%	0.1195	0.425	0.4	40.74%	55.22%	0.2391	0.2964	0.381	33.33%	50.00%	None	0.563445	0.4190476	34.47%	50.69%	0.0245	0.5634	0.419	37.04%	52.61%	0.1666	0.425	0.4
KNN	30.77%	47.96%	-0.088971294	0.695897937	0.438095238	31.91%	48.53%	-0.07799489	0.695897937	0.4381	41.88%	55.79%	0.151	0.2964	0.381	50.17%	62.48%	0.2985	0.0596	0.3238	56.98%	67.13%	0.3849	0.0127	0.2857	62.11%	71.21%	0.4621	0.001783	0.247619	64.67%	73.25%	0.5005	0.0006	0.2286	67.24%	75.28%	0.5392	0.0002	0.2095
Bagging	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419
Random Forest	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419
Ada Boost	33.33%	50.00%	None	0.563445001	0.419047619	35.90%	52.04%	0.149501678	0.424971477	0.4	34.47%	50.69%	0.0245	0.5634	0.419	41.88%	55.91%	0.1924	0.2964	0.381	44.16%	57.51%	0.1733	0.2964	0.381	48.15%	60.44%	0.3489	0.111445	0.3428571	39.32%	53.98%	0.1118	0.425	0.4	45.58%	58.51%	0.2542	0.19	0.3619
Linear SVC	33.90%	49.82%	-0.017348217	0.807973499	0.457142857	48.15%	60.44%	0.234729027	0.111445225	0.3429	50.71%	62.59%	0.2754	0.0596	0.3238	56.98%	67.24%	0.3793	0.0127	0.2857	50.71%	62.48%	0.3299	0.0596	0.3238	44.44%	57.83%	0.2333	0.190016	0.3619048	45.58%	58.51%	0.2169	0.19	0.3619	48.15%	60.44%	0.288	0.1114	0.3429
Passive Aggressive	37.61%	52.43%	0.031290312	0.695897937	0.438095238	48.15%	60.55%	0.2500464	0.111445225	0.3429	48.15%	60.67%	0.2382	0.1114	0.3429	50.71%	62.59%	0.2929	0.0596	0.3238	49.57%	61.91%	0.3729	0.0596	0.3238	50.71%	62.48%	0.3299	0.059555	0.3238095	39.32%	53.87%	0.1024	0.425	0.4	48.15%	60.44%	0.3489	0.1114	0.3429
Ridge	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	0	0.5634	0.419
Gradient Boosting	33.33%	50.00%	None	0.563445001	0.419047619	33.33%	50.00%	None	0.563445001	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	35.90%	52.04%	0.089	0.425	0.4
LDA	33.33%	50.00%	None	0.563445001	0.419047619	35.90%	52.04%	0.149501678	0.424971477	0.4	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.563445	0.4190476	33.33%	50.00%	None	0.5634	0.419	33.33%	50.00%	None	0.5634	0.419
QDA						41.88%	56.49%	0.13579302	0.296388919	0.381	61.54%	70.34%	0.401	0.0127	0.2857	50.43%	62.39%	0.2483	0.1114	0.3429	37.89%	53.56%	0.0783	0.5634	0.419	61.82%	70.89%	0.4075	0.005017	0.2666667	39.32%	54.10%	0.0762	0.425	0.4	42.17%	57.50%	0.171	0.19	0.3619
SGD	28.77%	45.74%	-0.104347907	0.945745329	0.495238095	50.71%	62.71%	0.28906592	0.05955485	0.3238	51.85%	63.28%	0.273	0.0596	0.3238	55.56%	66.01%	0.3151	0.0289	0.3048	50.71%	62.83%	0.3071	0.0596	0.3238	42.17%	56.69%	0.2742	0.190016	0.3619048	38.18%	53.30%	0.0937	0.425	0.4	45.58%	58.40%	0.2429	0.19	0.3619

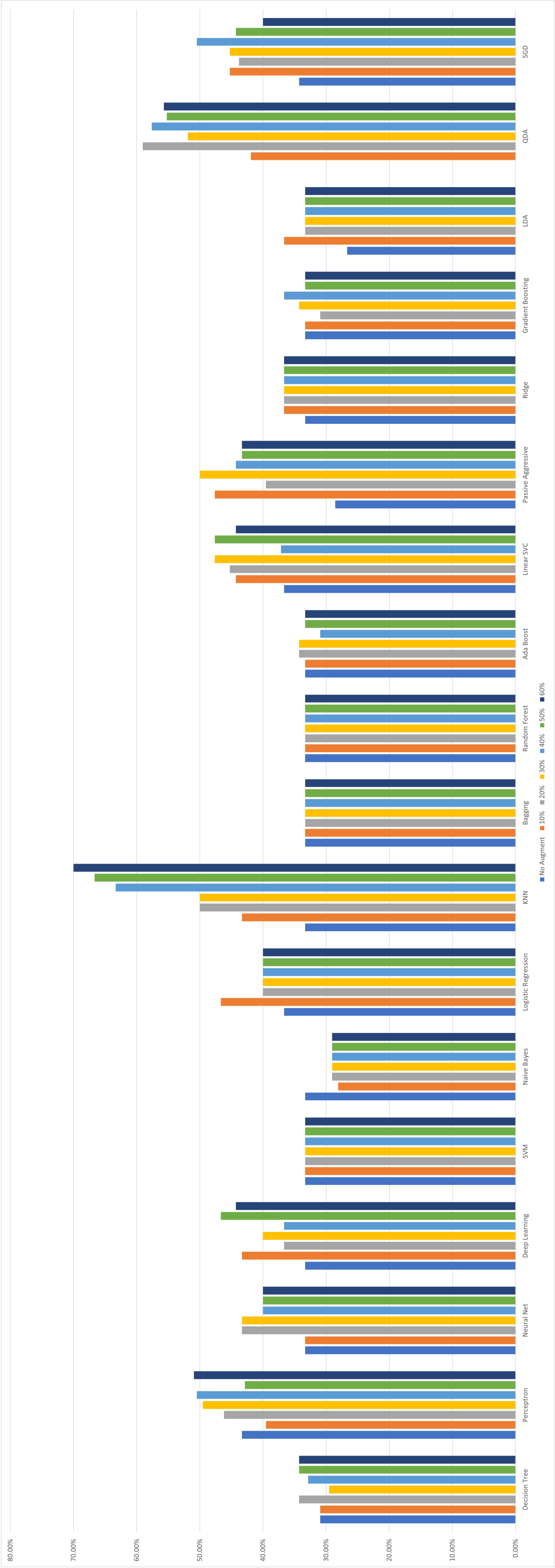


Figure D.1: *Openness trait prediction from gradually augmented dataset.*

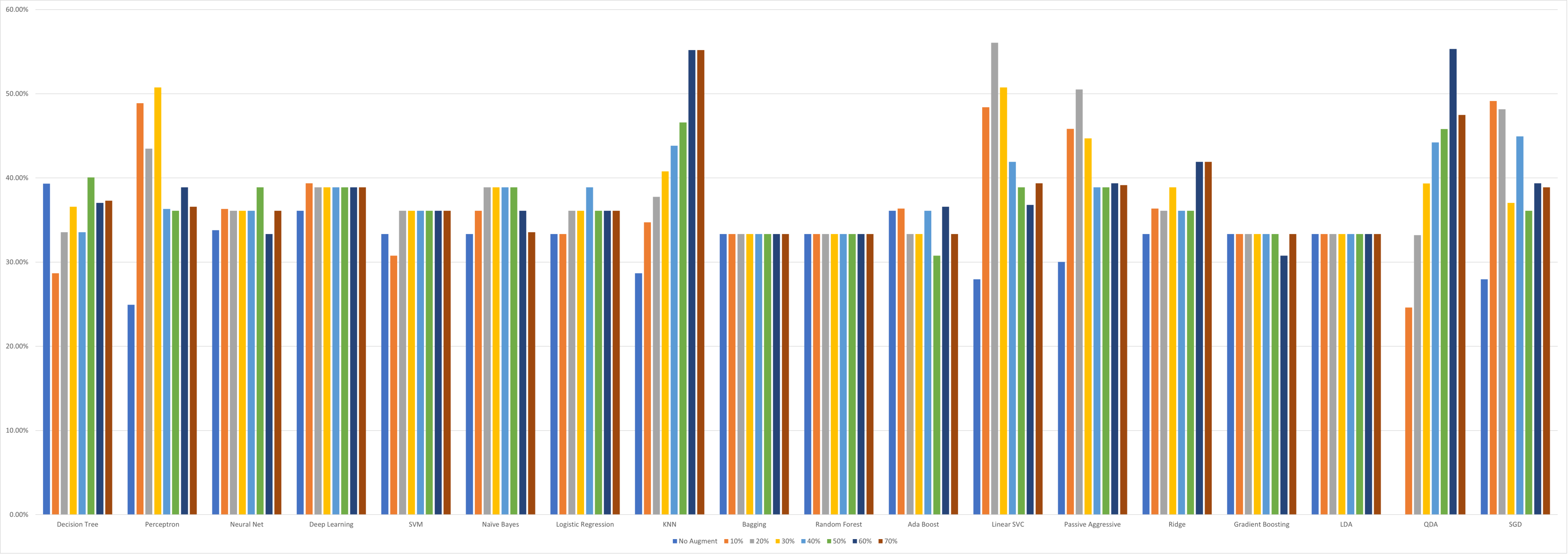


Figure D.2: *Conscientiousness trait prediction from gradually augmented dataset.*

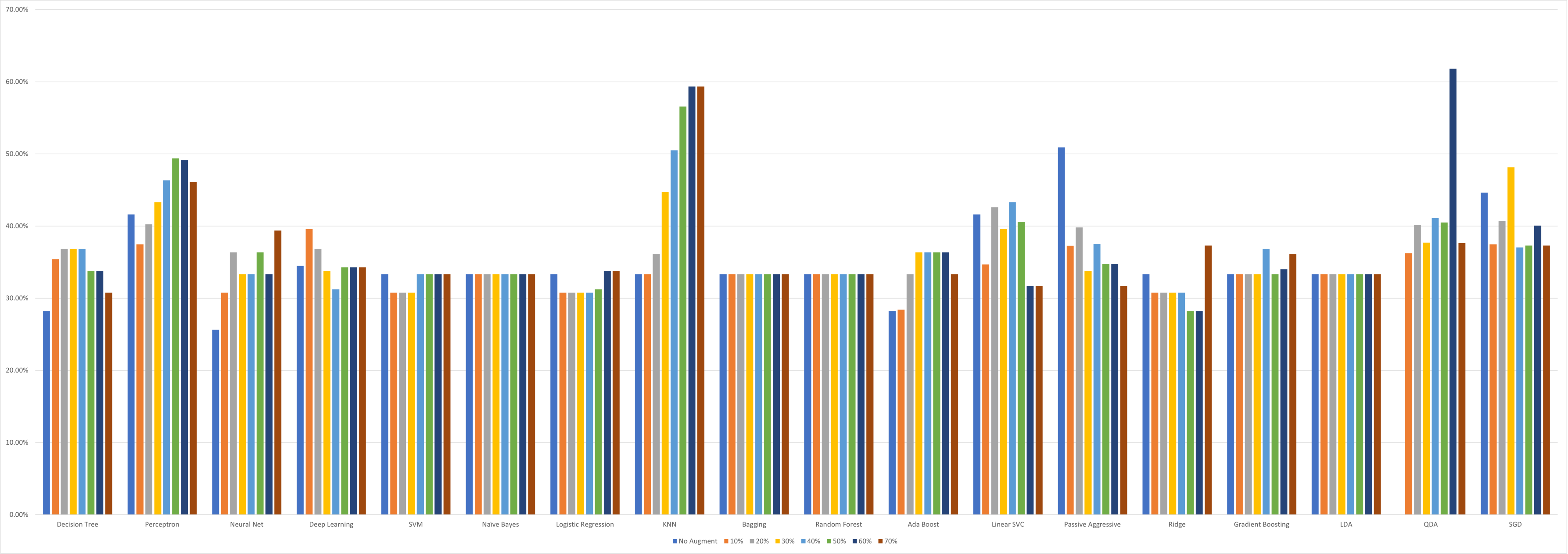


Figure D.3: *Extraversion trait prediction from gradually augmented dataset.*

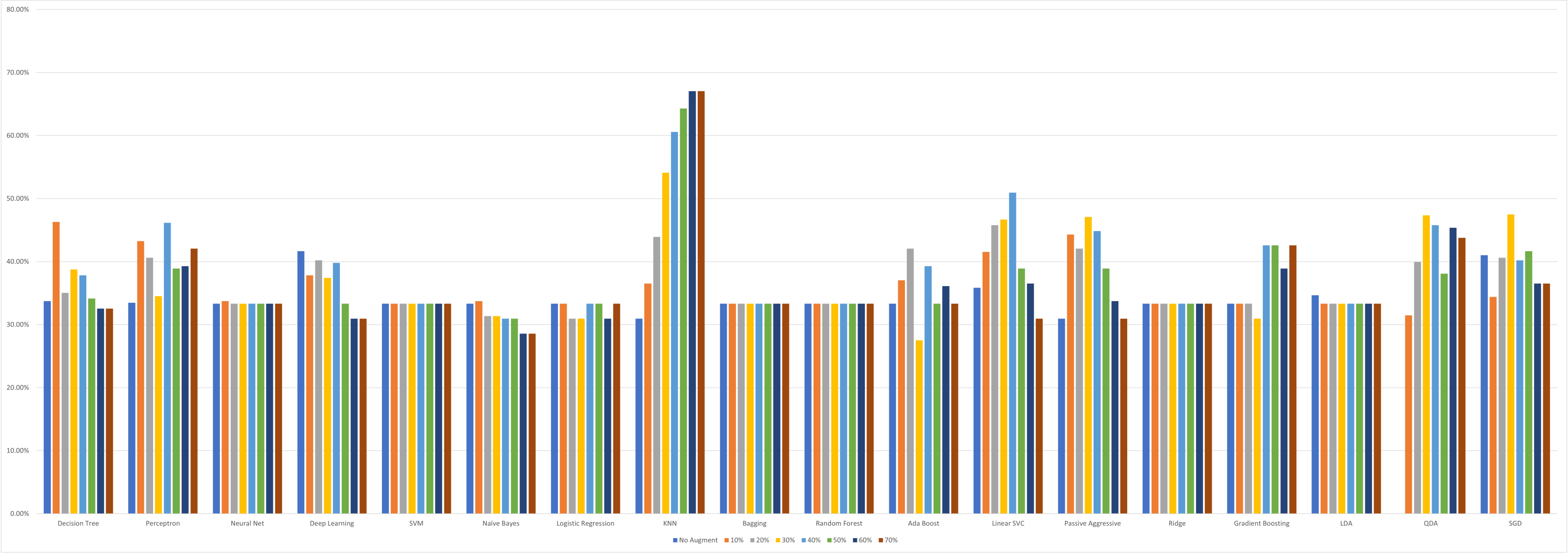


Figure D.4: *Agreeableness trait prediction from gradually augmented dataset.*

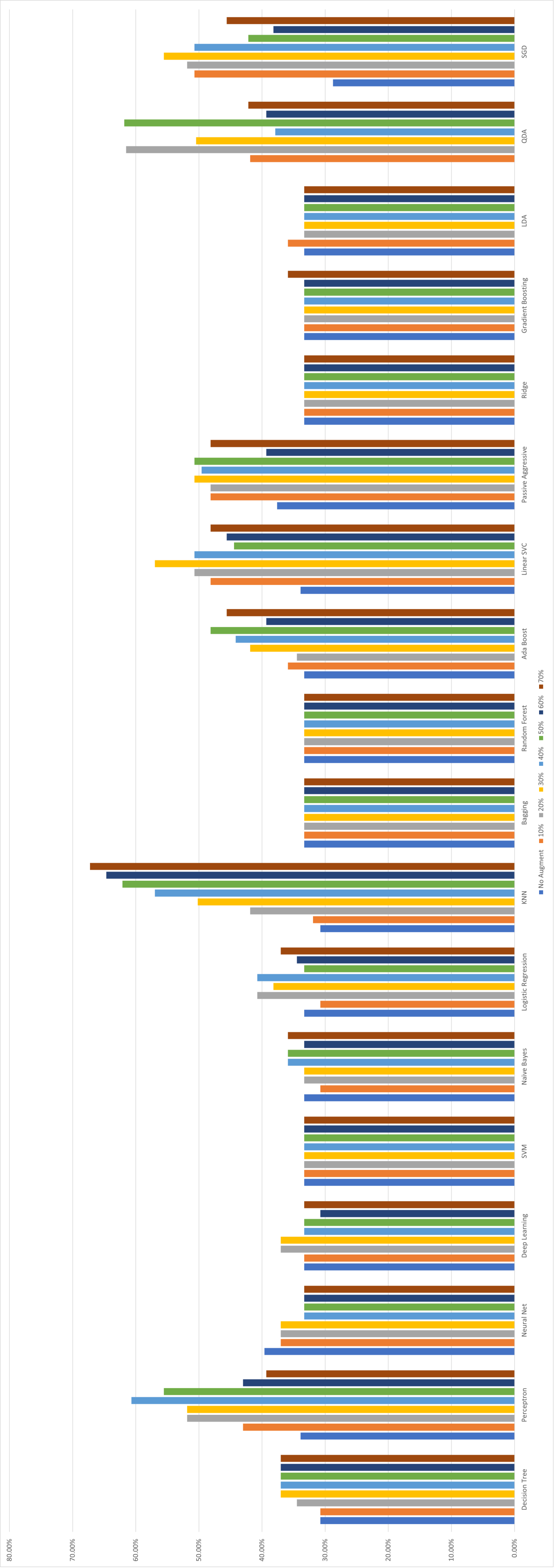


Figure D.5: *Neuroticism trait prediction from gradually augmented dataset.*

Appendix E: SPC Experiments: Hyperparameters Tuning

This appendix includes the tuned hyperparameters from the Speaker Personality Corpus experiments: 2, 5, and the modified experiment.

Table E.1: *Experiment 2: Accuracy Models with Hyperparameter Tuning - kNN*

kNN	Leaf Size	Metric	Neighbours	Weights
O	30	minkowski	28	Uniform
C	30	manhattan	21	Uniform
E	30	minkowski	1	Uniform
A	30	minkowski	4	Uniform
N	30	manhattan	6	Distance

Table E.2: *Experiment 2: Accuracy Models with Hyperparameter Tuning - SVM*

SVM	C	gamma	Kernel
O	1	0.5	rbf
C	1	0.5	rbf
E	1	0.5	rbf
A	1	scale	linear
N	10	0.001	rbf

Table E.3: *Experiment 2: Accuracy Models with Hyperparameter Tuning - LR*

LR	C	Max Iteration	Tol
O	0.0001	100	0.01
C	0.0001	100	0.01
E	1000	100	0.01
A	100	100	0.001
N	1000	100	0.0001

Table E.4: *Experiment 2: Accuracy Models with Hyperparameter Tuning - RF*

RF	Max Depth	Estimator
O	50	800
C	50	800
E	50	300
A	50	300
N	50	300

Table E.5: *Experiment 2: Accuracy Models with Hyperparameter Tuning - DL*

DL	Activation	Alpha	Hidden Layer	Max Iteration	Solver
O	relu	0.5	90	200	sgd
C	relu	0.5	100	200	sgd
E	relu	0.5	100	200	sgd
A	tanh	0.5	100	200	adam
N	relu	0.001	100	150	sgd

Table E.6: *Experiment 5: Recall Models with Hyperparameter Tuning - kNN*

kNN	Leaf Size	Metric	Neighbours	Weights
O	30	manhattan	30	Uniform
C	30	manhattan	10	Uniform
E	30	minkowski	1	Uniform
A	30	minkowski	4	Uniform
N	30	minkowski	16	Distance

Table E.7: *Experiment 5: Recall Models with Hyperparameter Tuning - SVM*

SVM	**C**	**gamma**	**Kernel**
O	1	scale	linear
C	1	0.5	rbf
E	1	0.5	rbf
A	1	scale	linear
N	1	scale	linear

Table E.8: *Experiment 5: Recall Models with Hyperparameter Tuning - LR*

LR	**C**	**Max Iteration**	**Tol**
O	100	100	0.01
C	25	100	0.01
E	10000	100	0.0001
A	25	100	0.001
N	10	100	0.001

Table E.9: *Experiment 5: Recall Models with Hyperparameter Tuning - RF*

RF	**Max Depth**	**Estimator**
O	50	1200
C	50	300
E	50	100
A	50	100
N	50	500

Table E.10: *Experiment 5: Recall Models with Hyperparameter Tuning - DL*

DL	**Activation**	**Alpha**	**Hidden Layer**	**Max Iteration**	**Solver**
O	tanh	0.5	200	200	sgd
C	tanh	0.001	200	200	sgd
E	relu	0.001	150	200	sgd
A	tanh	0.01	100	200	adam
N	relu	0.1	150	150	adam

Table E.11: *Modified Experiment: Hyperparameter Tuning - kNN*

kNN	Leaf Size	Metric	Neighbours	Weights
O	30	manhattan	14	Uniform
C	30	minkowski	22	Uniform
E	30	manhattan	2	Uniform
A	30	manhattan	1	Uniform
N	30	manhattan	3	Distance

Table E.12: *Modified Experiment: Hyperparameter Tuning - SVM*

SVM	C	gamma	Kernel
O	100	0.0001	rbf
C	1	scale	linear
E	100	0.0001	rbf
A	10	0.001	rbf
N	1	0.001	rbf

Table E.13: *Modified Experiment: Hyperparameter Tuning - LR*

LR	C	Max Iteration	Tol
O	0.09	100	0.01
C	0.09	100	0.01
E	0.009	100	0.01
A	100	100	0.01
N	1	100	0.01

Table E.14: *Modified Experiment: Hyperparameter Tuning - RF*

RF	Max Depth	Estimator
O	50	4
C	10	4
E	5	4
A	6	4
N	30	4

Appendix F: PTC Experiments: Hyperparameters Tuning

This appendix includes the tuned hyperparameters from the Personality Traits Corpus experiments 2, 3 and 4.

Table F.1: *Experiment 2: Hyperparameter Tuning - Openness Trait - No Augmentation*

Openness	**No Augmentation**
Decision Tree	criterion: entropy, max_leaf_nodes: 6, min_samples_split: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01
ANN	’activation’: ’tanh’, ’hidden_layer_sizes’: (), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.001, ’solver’: ’adam’
Deep Learning	’activation’: ’relu’, ’alpha’: 0.001, ’hidden_layer_sizes’: 90, ’max_iter’: 100, ’solver’: ’adam’
SVM	’C’: 1, ’kernel’: ’linear’
Naïve Bayes	’var_smoothing’: 1e-05
Logistic Regression	’C’: 5, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	n_neighbors=1, weights=uniform, leaf size=30, metric=manhattan
Bagging	’max_features’: 10, ’max_samples’: 50, ’n_estimators’: 800
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1.5, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: False, ’max_iter’: 10
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.1, ’n_estimators’: 100
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’hinge’, ’penalty’: ’elasticnet’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.2: *Experiment 2: Hyperparameter Tuning - Conscientiousness Trait - No Augmentation*

Conscientiousness	**No Augmentation**
Decision Tree	’criterion’: ’gini’, ’max_leaf_nodes’: 6, ’min_samples_split’: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.0001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.5, ’hidden_layer_sizes’: 70, ’max_iter’: 200, ’solver’: ’adam’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.01
Logistic Regression	’C’: 0.09, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 1, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 13, ’max_samples’: 50, ’n_estimators’: 800
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1.5, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 10
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.01, ’n_estimators’: 50
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.3: *Experiment 2: Hyperparameter Tuning - Extraversion Trait - No Augmentation*

Extraversion	No Augmentation
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 3, ’min_samples_split’: 5
Perceptron	’alpha’: 1e-05, ’penalty’: ’none’, ’tol’: 0.01
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: 5, ’learning_rate’: ’adaptive’, ’learning_rate_init’: 0.001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.001, ’hidden_layer_sizes’: 90, ’max_iter’: 100, ’solver’: ’sgd’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.1
Logistic Regression	’C’: 1, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.0001
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 4, ’weights’: ’distance’
Bagging	’max_features’: 10, ’max_samples’: 50, ’n_estimators’: 300
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 800
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.1, ’n_estimators’: 100
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.4: *Experiment 2: Hyperparameter Tuning - Agreeableness Trait - No Augmentation*

Agreeableness	No Augmentation
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 5, ’min_samples_split’: 5
Perceptron	’alpha’: 1e-05, ’penalty’: ’none’, ’tol’: 0.01
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.001, ’solver’: ’sgd’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.5, ’hidden_layer_sizes’: 70, ’max_iter’: 200, ’solver’: ’adam’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.1
Logistic Regression	’C’: 0.09, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 5, ’max_samples’: 50, ’n_estimators’: 500
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 100
Adaboost	’algorithm’: ’SAMME.R’, ’learning_rate’: 0.5, ’n_estimators’: 75
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.01, ’n_estimators’: 150
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’elasticnet’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.5: *Experiment 2: Hyperparameter Tuning - Neuroticism Trait - No Augmentation*

Neuroticism	No Augmentation
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 3, ’min_samples_split’: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01
ANN	’activation’: ’tanh’, ’hidden_layer_sizes’: (), ’learning_rate’: ’adaptive’, ’learning_rate_init’: 0.001, ’solver’: ’sgd’
Deep Learning	’activation’: ’relu’, ’alpha’: 0.001, ’hidden_layer_sizes’: 70, ’max_iter’: 100, ’solver’: ’sgd’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.01
Logistic Regression	’C’: 1, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 10, ’max_samples’: 25, ’n_estimators’: 300
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME.R’, ’learning_rate’: 0.5, ’n_estimators’: 100
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.001, ’n_estimators’: 200
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.001, ’loss’: ’log’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.6: *Experiment 3: Hyperparameter Tuning - Openness Trait - with Augmentation*

Openness	Augmentation
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 6, ’min_samples_split’: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01
ANN	’activation’: ’tanh’, ’hidden_layer_sizes’: (), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.001, ’solver’: ’adam’
Deep Learning	’activation’: ’relu’, ’alpha’: 0.001, ’hidden_layer_sizes’: 90, ’max_iter’: 100, ’solver’: ’adam’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 1e-05
Logistic Regression	’C’: 5, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 10, ’max_samples’: 50, ’n_estimators’: 800
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1.5, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: False, ’max_iter’: 10
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.1, ’n_estimators’: 100
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’hinge’, ’penalty’: ’elasticnet’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.7: *Experiment 3: Hyperparameter Tuning - Conscientiousness Trait - with Augmentation*

Conscientiousness	Augmentation
Decision Tree	’criterion’: ’gini’, ’max_leaf_nodes’: 6, ’min_samples_split’: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.0001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.5, ’hidden_layer_sizes’: 70, ’max_iter’: 200, ’solver’: ’adam’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.01
Logistic Regression	’C’: 0.09, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 1, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 13, ’max_samples’: 50, ’n_estimators’: 800
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1.5, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 10
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.01, ’n_estimators’: 50
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.8: *Experiment 3: Hyperparameter Tuning - Extraversion Trait - with Augmentation*

Extraversion	Augmentation
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 3, ’min_samples_split’: 5
Perceptron	’alpha’: 1e-05, ’penalty’: ’none’, ’tol’: 0.01
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: 5, ’learning_rate’: ’adaptive’, ’learning_rate_init’: 0.001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.001, ’hidden_layer_sizes’: 90, ’max_iter’: 100, ’solver’: ’sgd’
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.1
Logistic Regression	’C’: 1, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.0001
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 4, ’weights’: ’distance’
Bagging	’max_features’: 10, ’max_samples’: 50, ’n_estimators’: 300
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 800
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.1, ’n_estimators’: 100
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.9: *Experiment 3: Hyperparameter Tuning - Agreeableness Trait - with Augmentation*

Agreeableness	Augmentation
Decision Tree	'criterion': 'entropy', 'max_leaf_nodes': 5, 'min_samples_split': 5
Perceptron	'alpha': 1e-05, 'penalty': 'none', 'tol': 0.01
ANN	'activation': 'relu', 'hidden_layer_sizes': (), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'solver': 'sgd'
Deep Learning	'activation': 'tanh', 'alpha': 0.5, 'hidden_layer_sizes': 70, 'max_iter': 200, 'solver': 'adam'
SVM	'C': 1, 'kernel': 'linear'
Naive Bayes	'var_smoothing': 0.1
Logistic Regression	'C': 0.09, 'max_iter': 100, 'penalty': 'l2', 'tol': 0.01
kNN	'leaf_size': 30, 'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'
Bagging	'max_features': 5, 'max_samples': 50, 'n_estimators': 500
Random Forest	'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100
Adaboost	'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 75
Linear SVC	'C': 1
Passive Aggressive	'C': 1, 'fit_intercept': True, 'max_iter': 5
Ridge	'alpha': 0.1
Gradient Boosting	'learning_rate': 0.01, 'n_estimators': 150
LDA	'solver': 'svd', 'tol': 1e-05
SGD	'alpha': 0.0001, 'loss': 'squared_hinge', 'penalty': 'elasticnet', 'tol': 0.01
QDA	'tol': 1e-06

Table F.10: *Experiment 3: Hyperparameter Tuning - Neuroticism Trait - with Augmentation*

Neuroticism	10% Augmentation
Decision Tree	'criterion': 'entropy', 'max_leaf_nodes': 3, 'min_samples_split': 5
Perceptron	'alpha': 0.0001, 'penalty': 'l2', 'tol': 0.01
ANN	'activation': 'tanh', 'hidden_layer_sizes': (), 'learning_rate': 'adaptive', 'learning_rate_init': 0.001, 'solver': 'sgd'
Deep Learning	'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': 70, 'max_iter': 100, 'solver': 'sgd'
SVM	'C': 1, 'kernel': 'linear'
Naive Bayes	'var_smoothing': 0.01
Logistic Regression	'C': 1, 'max_iter': 100, 'penalty': 'l2', 'tol': 0.01
kNN	'leaf_size': 30, 'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'
Bagging	'max_features': 10, 'max_samples': 25, 'n_estimators': 300
Random Forest	'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300
Adaboost	'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 100
Linear SVC	'C': 1
Passive Aggressive	'C': 1, 'fit_intercept': True, 'max_iter': 5
Ridge	'alpha': 0.1
Gradient Boosting	'learning_rate': 0.001, 'n_estimators': 200
LDA	'solver': 'svd', 'tol': 1e-05
SGD	'alpha': 0.001, 'loss': 'log', 'penalty': 'l2', 'tol': 0.01
QDA	'tol': 1e-06

Table F.11: *Experiment 4: Hyperparameter Tuning - Openness Trait - Augmentation and Feature Reduction*

Openness	Augmentation and Feature Reduction
Decision Tree	'criterion': 'entropy', 'max_leaf_nodes': 5, 'min_samples_split': 5
Perceptron	'alpha': 0.001, 'penalty': 'l2', 'tol': 0.01
ANN	'activation': 'tanh', 'hidden_layer_sizes': (), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'solver': 'adam'
Deep Learning	'activation': 'tanh', 'alpha': 0.001, 'hidden_layer_sizes': 70, 'max_iter': 100, 'solver': 'adam'
SVM	'C': 1, 'kernel': 'linear'
Naive Bayes	'var_smoothing': 0.1
Logistic Regression	'C': 1, 'max_iter': 100, 'penalty': 'l2', 'tol': 0.01
kNN	'leaf_size': 30, 'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'
Bagging	'max_features': 2, 'max_samples': 50, 'n_estimators': 1200
Random Forest	'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100
Adaboost	'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 75
Linear SVC	'C': 1
Passive Aggressive	'C': 1, 'fit_intercept': True, 'max_iter': 10
Ridge	'alpha': 0.1
Gradient Boosting	'learning_rate': 0.01, 'n_estimators': 200
LDA	'solver': 'svd', 'tol': 1e-05
SGD	'alpha': 0.001, 'loss': 'hinge', 'penalty': 'l2', 'tol': 0.01
QDA	'tol': 1e-06

Table F.12: *Experiment 4: Hyperparameter Tuning - Conscientiousness Trait - Augmentation and Feature Reduction*

Conscientiousness	**Augmentation and Feature Reduction**
Decision Tree	’criterion’: ’gini’, ’max_leaf_nodes’: 6, ’min_samples_split’: 5
Perceptron	’alpha’: 0.0001, ’penalty’: ’l2’, ’tol’: 0.01, ’max_iteration’:1000
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (8,3), ’learning_rate’: ’constant’, ’learning_rate_init’: 0.0001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.5, ’hidden_layer_sizes’: 70, ’max_iter’: 200, ’solver’: ’adam’, learning_rate:0.004
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.01
Logistic Regression	’C’: 0.09, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 13, ’max_samples’: 50, ’n_estimators’: 800
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 5, ’min_samples_split’: 1, ’n_estimators’: 300
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1.5, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 10
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.01, ’n_estimators’: 50
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.13: *Experiment 4: Hyperparameter Tuning - Extraversion Trait - Augmentation and Feature Reduction*

Extraversion	**Augmentation and Feature Reduction**
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 3, ’min_samples_split’: 5
Perceptron	’alpha’: 1e-05, ’penalty’: ’none’, ’tol’: 0.01, ’max_iteration’:1000
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (5,5,5), ’learning_rate’: ’adaptive, ’learning_rate_init’: 0.001, ’solver’: ’adam’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.001, ’hidden_layer_sizes’: 70, ’max_iter’: 100, ’solver’: ’sgd’, learning_rate:0.001
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.1
Logistic Regression	’C’:1, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.0001
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 4, ’weights’: ’distance’
Bagging	’max_features’: 10, ’max_samples’: 1, ’n_estimators’:300
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 800
Adaboost	’algorithm’: ’SAMME’, ’learning_rate’: 1, ’n_estimators’: 150
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.1, ’n_estimators’: 100
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’l2’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.14: *Experiment 4: Hyperparameter Tuning - Agreeableness Trait - Augmentation and Feature Reduction*

Agreeableness	**Augmentation and Feature Reduction**
Decision Tree	’criterion’: ’entropy’, ’max_leaf_nodes’: 5, ’min_samples_split’: 5
Perceptron	’alpha’: 1e-05, ’penalty’: ’none’, ’tol’: 0.01, ’max_iteration’:1500
ANN	’activation’: ’relu’, ’hidden_layer_sizes’: (8,3), ’learning_rate’: ’constant, ’learning_rate_init’: 0.001, ’solver’: ’sgd’
Deep Learning	’activation’: ’tanh’, ’alpha’: 0.5, ’hidden_layer_sizes’: 70, ’max_iter’: 200, ’solver’: ’adam’, learning_rate:0.001
SVM	’C’: 1, ’kernel’: ’linear’
Naive Bayes	’var_smoothing’: 0.1
Logistic Regression	’C’:0.09, ’max_iter’: 100, ’penalty’: ’l2’, ’tol’: 0.01
kNN	’leaf_size’: 30, ’metric’: ’manhattan’, ’n_neighbors’: 1, ’weights’: ’uniform’
Bagging	’max_features’: 5, ’max_samples’: 1, ’n_estimators’:500
Random Forest	’max_depth’: 50, ’min_samples_leaf’: 1, ’min_samples_split’: 2, ’n_estimators’: 100
Adaboost	’algorithm’: ’SAMME.R’, ’learning_rate’: 0.5, ’n_estimators’: 75
Linear SVC	’C’: 1
Passive Aggressive	’C’: 1, ’fit_intercept’: True, ’max_iter’: 5
Ridge	’alpha’: 0.1
Gradient Boosting	’learning_rate’: 0.01, ’n_estimators’: 150
LDA	’solver’: ’svd’, ’tol’: 1e-05
SGD	’alpha’: 0.0001, ’loss’: ’squared_hinge’, ’penalty’: ’elasticnet’, ’tol’: 0.01
QDA	’tol’: 1e-06

Table F.15: *Experiment 4: Hyperparameter Tuning - Neuroticism Trait - Augmentation and Feature Reduction*

Neuroticism	**20% Augmentation**
Decision Tree	'criterion': 'entropy', 'max_leaf_nodes': 6, 'min_samples_split': 5
Perceptron	'alpha': 0.0001, 'penalty': 'l2', 'tol': 0.01
ANN	'activation': 'tanh', 'hidden_layer_sizes': (), 'learning_rate': 'adaptive', 'learning_rate_init': 0.001, 'solver': 'sgd'
Deep Learning	'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': 70, 'max_iter': 100, 'solver': 'sgd'
SVM	'C': 1, 'kernel': 'linear'
Naive Bayes	'var_smoothing': 0.01
Logistic Regression	'C': 1, 'max_iter': 100, 'penalty': 'l2', 'tol': 0.01
kNN	'leaf_size': 30, 'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'
Bagging	'max_features': 10, 'max_samples': 25, 'n_estimators': 300
Random Forest	'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300
Adaboost	'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 100
Linear SVC	'C': 1
Passive Aggressive	'C': 1, 'fit_intercept': True, 'max_iter': 5
Ridge	'alpha': 0.1
Gradient Boosting	'learning_rate': 0.001, 'n_estimators': 200
LDA	'solver': 'svd', 'tol': 1e-05
SGD	'alpha': 0.001, 'loss': 'log', 'penalty': 'l2', 'tol': 0.01
QDA	'tol': 1e-06