Leveraging Custom CNN and ResNet Architectures for Identifying Deviant Personality Traits in Cyber Threat Detection Using the Five Factor Model

Iustin FLOROIU
National Institute for Research and Development in Informatics – ICI Bucharest
iustin.floroiu@ici.ro

Abstract: Understanding personality factors play a crucial role in the modelling of individuals' behavioural patterns with regard to specific risk factors. A major problem in psychology related to this aspect revolves around using different computational algorithms and models to automatically verify the risk of cyber threats in both online and physical environments. The major attempt this paper is trying to make is detecting different disorderly behaviours that play a crucial role in risk management with regard to individuals, especially in the context of cyberterrorism. Custom convolutional and residual networks were implemented as the key solution to solving this contextual problem.
Keywords: cybersecurity, convolutional networks, residual networks, big5.

INTRODUCTION

In the digital age, cybersecurity remains a paramount concern, with threats continuously evolving in sophistication. While technical defences against cyberattacks have improved, there is a growing need to understand the human factor—specifically, the personality traits of individuals who engage in cyberthreats. The Five Factor Model (FFM), or Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), offers a well-established framework for studying human behaviour and personality (Wiggins, 1996). Understanding how these traits influence risky online behaviours can provide valuable insights into the psychological predispositions underlying cybercriminal activities with the aid of artificial intelligence algorithms that have greatly advanced in fields such as cybersecurity (Floroiu, 2024), medical applications (Paraschiv, 2024), (Floroiu, 2024) and in many more fields.

https://doi.org/10.54851/v6i2y202406

Recent studies indicate significant correlations between certain FFM dimensions and behaviours relevant to cybersecurity. For instance, low Agreeableness and Conscientiousness have been associated with a higher likelihood of engaging in cybercrime and unethical online behaviour (Sudzina, 2020), while high Neuroticism may increase vulnerability to social engineering attacks (de Weijer, 2017). These findings suggest that personality traits influence not only cyberthreats, but also broader behaviours, such as religious orientation and dutifulness, influenced by Openness and Conscientiousness, respectively (Silvia, 2020; Wilmot, 2019).Despite this growing body of research, a significant gap remains in how these psychological insights can be systematically integrated into cybersecurity strategies. Specifically, the application of deep learning models to analyse personality traits in relation to cyberthreats has not been fully explored. Current models primarily focus on technical indicators of attacks, often overlooking the psychological drivers behind them. Eventhough several transformer models can be approached, because of the vast contributions to the field (Floroiu, 2024), choosing convolutional and residual networks was based on the scarcity of samples found within the dataset.

THEORETICAL NOTIONS ON THE FIVE FACTOR PERSONALITY TRAITS MODEL

The Five Factor Model (FFM) of personality, also known as the Big Five personality traits, represents a comprehensive framework for understanding and categorising human personality. Rooted in decades of psychological research, the FFM posits that personality can be described along five broad dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Each of these dimensions captures distinct aspects of an individual's behavioural tendencies, emotional patterns, and cognitive styles, offering valuable insights into the complexities of human nature.At the core of the FFM lies the belief that personality traits are relatively stable and enduring across different situations and contexts, shaping how individuals perceive the world, interact with others, and navigate life's challenges. While individuals may vary in the degree to which they express each trait, these dimensions provide a useful framework for understanding the fundamental building blocks of personality and predicting behaviour across diverse situations.

Openness to Experience reflects the extent to which individuals are receptive to new ideas, experiences, and perspectives. Those high in openness tend to be imaginative, curious, and open-minded, eagerly exploring novel concepts and embracing unconventional viewpoints. They are often drawn to artistic endeavours, intellectual pursuits, and cultural experiences, seeking to expand their horizons and deepen their understanding of the world (Abood, 2019).

Conscientiousness encompasses traits related to self-discipline, organisation, and goal-directed behaviour. Individuals high in conscientiousness are diligent, responsible, and methodical in their approach to tasks and responsibilities. They set high standards for themselves, strive for excellence, and demonstrate a strong sense of reliability and accountability in their actions (Abood, 2019).

Extraversion captures the extent to which individuals are outgoing, sociable, and energized by social interactions. Extraverts thrive in social settings, enjoying the company of others and seeking out opportunities for social engagement and stimulation. They are often described as outgoing, assertive, and enthusiastic, with a natural ability to connect with people and cultivate social networks (Abood, 2019).

Individuals high in agreeableness are warm, altruistic, and considerate of others' needs and feelings. They value harmonious relationships and strive to maintain peace and goodwill in their interactions with others, often acting as empathetic listeners and supportive companions (Abood, 2019).

Neuroticism encompasses traits related to emotional stability, resilience, and susceptibility to negative emotions. Individuals high in

neuroticism are prone to experiencing feelings of anxiety, insecurity, and emotional distress in response to stressors and life challenges.

They may exhibit tendencies towards worry, self-doubt, and moodiness, often struggling to maintain emotional equilibrium in the face of adversity (Abood, 2019).

THEORETICAL NOTIONS ON THE DEEP LEARNING ARCHITECTURES IMPLEMENTED

Convolutional Networks

Convolutional Neural Networks (CNNs) are deep learning architectures that have fundamentally transformed the field of computer vision owing to their superior efficiency in processing pixel data. Over the years, their utility has transcended image processing, reaching into various domains, including natural language processing (NLP), audio recognition, and beyond. At their core, CNNs leverage the mathematical operation known as convolution, replacing general matrix multiplication in at least one of their layers (Yin, 2017).

A typical CNN architecture comprises multiple layers that can be grouped into three primary categories: convolutional layers, pooling layers, and fully connected layers. Each type of layer serves a unique function. Convolutional layers are the core building blocks of a CNN. They apply a convolution operation to the input, passing the result to the next layer. The convolution emulates a sliding window over the input data, extracting features from these local regions through learned kernels (filters). This mechanism allows CNNs to achieve translational invariance, making them very effective in recognising patterns anywhere in the input space.Following convolutional layers, pooling layers serve to reduce the spatial size of the convolved features. This downsampling is critical not only for lowering computational overhead but also for achieving an abstraction of the feature maps. Max pooling and average pooling are the two most common types.Towards the end, CNNs typically contain one or more fully connected layers where all neurons from the previous layer are connected to every neuron on the current layer. These layers are crucial for combining the features learned by the network to make predictions or classifications.Though initially designed for image data(Paraschiv, 2024), (Alassiri, 2022), CNNs have been successfully adapted for NLP tasks(Feng, 2023), (Gavrila, 2021). Their ability to extract and learn patterns from spatial data has been exploited to capture the contextual relationships in text. Efficient training of CNNs typically utilises backpropagation and optimisation algorithms such as Adam or RMSprop, which adapt the learning rate for each parameter. Regularisation techniques like dropout and batch normalisation are also widely used to prevent overfitting and ensure the generalization of the model to new data.

Residual Networks

Residual Networks (ResNets) are a class of deep learning models designed to enable the training of substantially deeper neural networks than was previously feasible. Introduced by He et al. in their seminal 2015 paper, ResNets address the vanishing/exploding gradient problem that often occurs with increasing network depth, thereby facilitating the training of networks that are hundreds, or even thousands, of layers deep. This is achieved through the novel use of „residual blocks" with skip connections that allow for alternate shortcut paths for gradient flow during backpropagation (Yin, 2017).The defining feature of ResNets is the residual block, where the main idea is to introduce a shortcut that skips one or more layers. Typical residual blocks contain two main paths. The main path consists of the sequential convolutional layers which typically follow the structure of Batch Normalization (BN) and then a Rectified Linear Unit (ReLU) activation function preceding a convolution layer. The shortcut path provides a shortcut for the gradient, allowing the signal to bypass some of these layers. The shortcut commonly involves identity mapping, where the input is added directly to the output of

the convolutional layers. If the dimensions differ, a linear projection is used to match the dimensions.These blocks help preserve the gradient from input to output, thereby allowing very deep networks to remain trainable.The key functionality of ResNets lies in tackling the degradation problem: As networks grow deeper, accuracy gets saturated and then quickly degrades. Historically, deeper networks were thought to be harder to train due to vanishing gradients, where the gradients of the loss function approach zero as they propagate back through the deep layers. However, the skip connections in residual networks mitigate this effect by allowing the gradient to flow through the network unimpeded. Layers in deep residual networks theoretically learn incremental changes (residuals) to the identity mappings of previous layers rather than complete transformations from inputs to outputs, making it easier for the model to converge.Though originally developed for image recognition tasks, ResNets have found vast applications across various domains. The training of ResNets, whilst similar to other deep neural networks involving forward and backward propagations, typically enjoys better convergence rates thanks to effective gradient flow. By enabling the training of significantly deeper networks without the risk of vanishing gradients, ResNets have broadened the horizon of what can be achieved with deep learning across multiple disciplines.

DATASET DESCRIPTION

The Essays Dataset is primarily derived from the myPersonality project, initiated by David Stillwell and Michal Kosinski (Kosinski, 2013). The myPersonality project aimed to collect and analyses ocial media data to study psychological traits. The specific dataset used in this research, often referred to as the „Essays Dataset," includes written essays from participants who also completed a Big Five personality assessment (Kosinski, 2014).The Essays Dataset is structured into two main components: the text data (essays) and the personality traits scores. The dataset includes the following fields: Essay_ID (a unique identifier for each essay), Essay_Text (the content of the essay written by the participants) and the big5 trait scores (5 scores, each indicating the scores for each personality trait, using binary labels based on the score for each trait being above or below average). The essays vary in length, style, and subject matter, reflecting the diversity of the participants' backgrounds and writing abilities. The dataset includes a wide range of topics, from personal reflections and life experiences to opinions on various subjects. This diversity makes the dataset particularly valuable for studying the relationship between language use and personality traits (Majumdar, 2023). Moreover, this dataset was chosen because of the scarcity of existent datasets for five factor model trait detection in essays. Another important factor in choosing this dataset regards the necessity of human-generated data for better training quality, even though synthetic generated datasets of FFM trait labels in essays exist (Floroiu, 2024).

TECHNICAL IMPLEMENTATION OF THE ALGORITHMS

The initial phase involves the preprocessing of the dataset comprised of state-of-consciousness essays. Each essay is systematically converted into a machine-readable format through tokenisation, where the text is split into individual words or phrases. Subsequently, these tokens are transformed into sequences of integers using a tokenizer fit on the corpus, ensuring each token is uniquely identifiable.

This step involves analysing the linguistic and psychological indicators present in the essays that correlate with specific personality dimensions outlined in the Five Factor Model (Big Five) - Neuroticism, Agreeableness, Extraversion, Conscientiousness, and Openness to Experience. Based on these traits, a function classify_personality is defined to categorise each essay into distinct personality classes, such as Narcissistic Personality, Borderline Traits, and Antisocial Traits, depending on the presence or absence of certain trait indicators

within the essay.Each essay in the dataset is encoded with binary indicators for each of the five major personality traits. These indicators (‚y' for yes, ‚n' for no) signify whether a particular trait is strongly exhibited within the essay according to predefined psychological assessment criteria. The function classify_ personality maps these binary combinations to a specific numeric label, ensuring that the subsequent model treats these as distinct classes.The classify_personality function is applied across the dataset using a DataFrame's apply method, which efficiently handles the row-wise application of complex mappings. The function checks each row for certain combinations of traits (encoded as binary) and assigns a numerical label according to the criteria set for identifying specific personality profiles.Once all essays in the dataset have been assigned preliminary numeric labels based on their personality categorisation, these labels are further processed using a label encoder. This transformation converts the numeric labels into a form suitable for the model's output layer, which predicts these categories as different classes.Given the potential for imbalances in class distribution (some personality types may be less common in the dataset), it is crucial to compute class weights. These weights adjust the importance given to each class during model training, compensating for underrepresented personality types by emphasising their significance in the loss function computation. Finally, after transforming text data into sequences and mapping personality traits to numeric classes, the dataset is divided into features (padded sequences) and targets (encoded labels). These are used as inputs to the deep learning models during the training and validation phases.

To simplify the problem, antisocial traits were correlated with low conscientiousness, openness, neuroticism and agreeableness and high extraversion; borderline traits were understood as low extraversion, conscientiousness, agreeableness, neuroticism and high openness to experience, while narcissism was correlated with low conscientiousness, agreeableness, openness to experience and high extraversion and neuroticism. While individual differences might exist, these trait patterns were considered the most influential in the detection of personality behaviours that might be seen as dangerous.To manage varying lengths of essays, padding is applied to standardize the sequence lengths, enabling uniform input sizes for the neural networks. Additionally, label encoding is employed to convert categorical labels into a numerical format suitable for model training, optimising the classification process.These embeddings provide a dense representation in which words with similar meanings are mapped to proximate points in the embedding space. The utilisation of embeddings Word2Vec allows the models to leverage learned word associations from vast amounts of text data, enhancing the model's ability to understand and process the input essays effectively (Mikolov, 2013).

A custom convolutional neural network model is tailored specifically for text analysis. The model architecture features multiple convolutional layers that apply various filters to the embedded word sequences, capturing different textual patterns. These filters help to identify essential features in the essays without the need for manual feature engineering. Following convolutional operations, global max-pooling layers are used to condense the feature maps into a single vector per map, emphasising the most salient features extracted from each map. The pooling layer significantly reduces the dimensionality of the data, which simplifies the network structure and reduces computation. Dense layers follow the pooling layers, serving as fully connected layers that perform high-level reasoning based on the features extracted previously. Dropout layers are interspersed among the dense layers to prevent overfitting by randomly dropping units during the training process, thereby enhancing the model's generalisation capabilities.In addition to the CNN model, a custom residual network architecture is developed to handle deeper layers without falling prey to the vanishing gradient problem. The ResNet model introduces skip connections that bypass one or more layers

and add the input directly to a deeper layer. This configuration allows gradients to flow through the network more effectively during training, facilitating the use of much deeper network architectures. The residual blocks consist of convolutions followed by batch normalisation and activation functions, with skip connections linking the input of the block to its output. This setup fosters learning residual functions with reference to the layer inputs, enabling training on residual mappings rather than unreferenced functions.Both models are compiled with an Adam optimizer, a popular choice for its adaptive learning rate capabilities, making it well-suited for datasets with sparse gradients like text data. The loss function used is sparse categorical cross-entropy, ideal for multi-class classification problems. Class weights are computed to address class imbalance within the training dataset, ensuring that the model does not develop a bias toward the more frequent classes. These weights are applied during training to modify the loss function, emphasising the importance of correctly predicting under-represented classes.

For the Convolutional Neural Network (CNN) model, the architecture commences with an Embedding layer, which is initialised with pre-trained Word2Vec embeddings and is set to non-trainable to retain the semantic properties learned from extensive external corpora. Following the embedding, the CNN model features two Conv1D layers; the first has 128 filters with a kernel size of 5, and the second also has 128 filters but with a kernel of size 7, both utilising ‚relu' activation. These convolutional layers are designed to extract local feature mappings from the sequences. Post convolution, a GlobalMaxPooling1D layer is applied to reduce the spatial dimensions by taking the maximum value over the time dimension for each feature, simplifying the network by abstracting the most significant features from the convolutions. Towards the output, the model contains a Dense layer with 128 units and ‚relu' activation, followed by a Dropout layer set at 0.5 to mitigate overfitting by randomly ignoring a subset of features during training. The final output layer is a Dense layer with 4 units (corresponding to the number of personality classes) and employs ‚softmax' activation for multi-class classification.

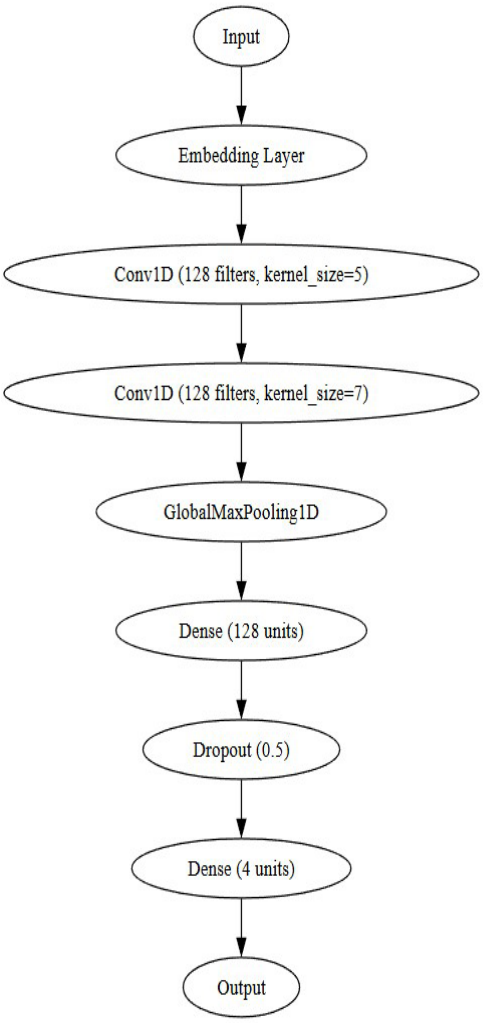

Figure 1. *Progress of ImageNet Classification Error from Feature Engineering to Deep Learning (2010-2017)*

The Residual Network (ResNet) model is similarly initiated with an Embedding layer with the same configuration as the CNN model for consistent input processing. The significant difference lies in the use of residual blocks, which incorporate two Conv1D layers each, all equipped with ‚relu' activation and batch normalisation. Each block features a skip connection that adds the input of the block directly to its output, facilitating the training of deeper networks by improving gradient flow.

Typically, each Conv1D within the residual block has 64 filters and a kernel size of 5, mirrored across each block to maintain architectural consistency. Following the residual blocks, a GlobalMaxPooling1D layer is employed to distil the features down to a manageable size. The classifier head includes a Dense layer with 128 units (with ‚relu' activation), a subsequent Dropout layer at a rate of 0.5, and concludes with a final Dense layer with 4 output units featuring ‚softmax' activation. This architecture leverages deep feature extraction without the learning impediments posed by increased depth, thanks to the residual connections.

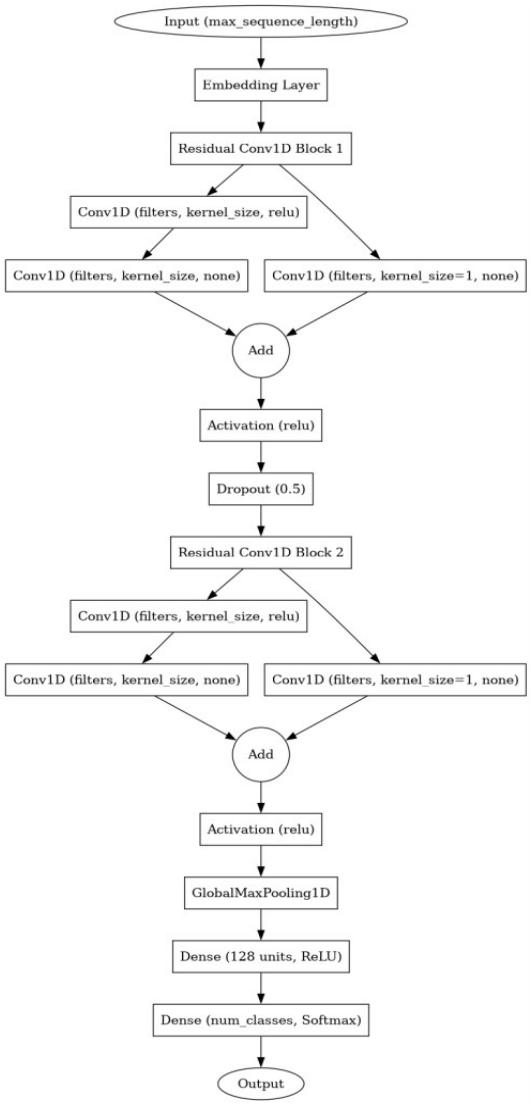

Figure 2. *The custom ResNet architecture*

Both models are structured to capitalise on their respective architectural strengths, with the CNN excelling in extracting hierarchical features through layered convolutions and the ResNet mitigating the challenges of training deeper networks via residual learning. These choices in layers, activations, and connections are specifically designed to tackle the nuances of textual data embedded within the state-of-consciousness essays, targeting an effective interpretation of underlying personality traits.

RESULTS

The custom Convolutional Neural Network (CNN) developed for analysing state-of-consciousness essays and classifying personality traits has undergone evaluation over 20 training epochs. This chapter discusses the outcomes of this training, as evidenced by alterations in accuracy and loss metrics across both the training and validation datasets. Graphical representations of training and validation accuracy, alongside training and validation loss, provide a visual narrative of the model's performance over time.

Throughout the training process, various key observations were made. The model began with a highly fluctuating start in both accuracy and loss. In the first epoch, the training accuracy opened at 42.05%, but the validation accuracy was markedly low at 4.86%. This discrepancy suggests that the model excessively learned the detailed noise of the training data at the cost of its generalisation ability on unseen data. As training progressed through epochs 2 to 10, a notable improvement in validation accuracy and a steady decrease in validation loss were observed. By the 10th epoch, the training accuracy rose substantially to 95.48%, and validation accuracy improved impressively to 90%, indicative of the model adjusting and better generalising from the learning features. From epoch 11 onward, both training and validation accuracy levels began to stabilise, with training accuracy peaking at 99.8% during epoch 17, indicating near-perfect learning on the training set.

Validation accuracy reached a plateau of around 90%, reflecting a strong but slightly limited ability to generalise the learned patterns to new data.

Figure 3. *The training and validation accuracy of the custom CNN model*

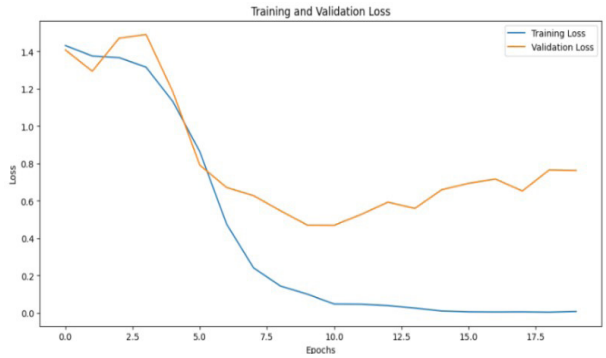

Figure 4. *The training and validation loss of the custom CNN model*

The training and validation accuracy graph shows a convergence point around epoch 10, beyond which improvements in training do not translate into better validation performance. This pattern is typical in neural network training cycles, where initial leaps in learning are followed by a plateau as the model exhausts the available generalisable features.Conversely, the training and validation loss graph provides a mirror image, with validation loss decreasing initially and then slowly creeping upwards as the model begins to overfit to the training data. The divergence between training loss and validation loss post-epoch 10 further substantiates a possible onset of overfitting. Overall, the results affirm the potential of the custom CNN architecture in processing and classifying complex textual data based on intrinsic personality traits, with key insights gained on optimising model performance for future iterations.

The custom Residual Network (ResNet) model designed to handle state-of-consciousness essays for personality identification has been put through a comprehensive evaluation across 20 epochs.Visual representations in the form of graphs supplement the analysis, providing a clear view of trends in model learning and validation dynamics.The focus is on pivotal metrics that demonstrate learning effectiveness—accuracy and loss—both of which provide insights into how well the model not only captures patterns in the training data but also generalises to unseen data.

Figure 5. *The training and validation accuracy of the custom ResNet model*

Figure 6. *The training and validation accuracy of the custom ResNet model*

The training commenced with the model showing a low initial accuracy of approximately 23.72%, coupled with a relatively high initial loss of 2.5932. On the validation front, the initial accuracy was near the equivalent of 24.09% with a loss of 1.3845. This starting point signaled substantial room for improvement, as the model had only begun to explore the patterns within the training dataset.From the second epoch, a discernible improvement was evident. The model's accuracy on training data enhanced to 32.39%, albeit with a surprising dip in validation accuracy to 9.92%. This fluctuation raised initial concerns about potential overfitting—where the model could overly adapt to the training data specifics at the cost of its general applicability. As training progressed towards the mid epochs, from the third through to the tenth epoch, both training and validation accuracies displayed a consistent upward trajectory. Concretely, the training accuracy improved robustly from 40.74% in the third epoch to 92.05% in the tenth epoch. Importantly, this was not at the expense of validation accuracy, which also increased significantly, peaking at 89.47% in the tenth epoch.The corresponding loss metrics mirrored these improvements. Training loss steadily declined, reaching a low of 0.1330 by the tenth epoch. Validation loss, which tends to be more volatile but is an equally crucial indicator of model generalisation, also decreased consistently to 0.4846 by the tenth epoch. This simultaneous decrease in loss alongside an increase in accuracy indicated effective learning and adaptation by the model without the drawback of memorising the training data—a common pitfall known as overfitting.The later epochs, from the eleventh to the twentieth, showed an interesting pattern. While the training accuracy continued to climb, peaking at around 98.96% in the fifteenth epoch, validation accuracy began to exhibit signs of plateauing. This was particularly noticeable around epochs 12 through 20, with validation accuracy fluctuating modestly within the 88.66% to 90.49% range. This stagnation could suggest the beginning of overfitting despite the various regularization efforts potentially in place. Similarly, training loss continued to decrease, reaching an optimal low of 0.0455 by the fifteenth epoch but increasing slightly afterwards. Validation loss trends displayed minor peaks and troughs, generally maintaining a range between 0.4473 and 0.5604.

These fluctuations are indicative of the model's attempt to generalise findings while being slightly perturbed by new patterns or noise in the validation set.The graphical representations of training and validation accuracies and losses provide a visual confirmation of the numerical data.The pattern of results indicates a robust learning process during the training of the custom ResNet model with excellent potential based on consistently high training accuracies and decreasing losses. However, the signs of early plateauing in validation accuracy accompanied by minor fluctuations in validation loss recommend caution. Overall, the trajectory suggests a promising model but underscores the continuous need for monitoring and tweaking to achieve the best balance between fitting and generalising.

CONCLUSIONS

The findings of this study demonstrate the significant potential of custom deep learning models, particularly the tailored ResNet, in detecting and analysing complex personality traits such as antisocial, narcissistic, and borderline characteristics. These traits have been closely linked to risky online behaviours and cyber threats, underscoring the importance of integrating psychological insights into cybersecurity strategies. Although the model exhibited excellent learning indicators during the early stages of training, signs of overfitting emerged in the later stages, highlighting the need for further optimisation. Nonetheless, the model performed well in capturing patterns from the training data, which lays a solid foundation for its application in real-world cybersecurity contexts.This research answers the key questions posed at the outset by demonstrating how personality traits from the Five Factor Model (FFM) can be harnessed to detect cybersecurity risks.

The study reveals that traits such as low Agreeableness and Conscientiousness, alongside high Neuroticism, correlate with a higher likelihood of engaging in cybercrime or falling victim to social engineering attacks. By applying deep learning techniques to analyse these traits, the study shows that psychological profiling can be a powerful tool in predicting cyber threats and identifying individuals who may be predisposed to malicious online behaviour.

In addition, this research provides evidence that deep learning models, such as convolutional neural networks (CNNs) and residual networks (ResNets), are highly effective in analysing sequential personality data. These models can reveal subtle patterns that may not be immediately apparent in traditional psychological assessments. For example, ResNets have proven especially useful in capturing deep and nuanced patterns from personality data without losing key information, even as the model's depth increases. This capacity for granular analysis suggests that such models could be invaluable in enhancing current cybersecurity protocols by identifying behavioural vulnerabilities before they are exploited.

The broader implications of this research extend beyond simply improving the accuracy of cyber threat detection. By integrating psychological profiling with security protocols, organisations can develop more targeted and proactive defence mechanisms. For instance, the personality traits identified by these models could inform personalised cybersecurity training programs, helping individuals become more resilient to cyber manipulation based on their psychological tendencies. Furthermore, in corporate environments, personality profiling could be used to design more effective access control systems, where employees with traits linked to higher cybersecurity risks might face stricter monitoring or restrictions. This could dramatically reduce the incidence of insider threats, which remain a significant challenge for many organisations.

Moreover, the use of deep learning models to analyse personality traits opens up new possibilities in user authentication systems. Behavioural patterns could be incorporated into multi-factor authentication protocols, enhancing security by accounting not only for physical credentials but also for the psychological tendencies of users. Such systems could make it more difficult for attackers to exploit human vulnerabilities, creating a more robust defence against social engineering and other manipulation-based cyberattacks.

While the integration of psychological profiling into cybersecurity measures offers exciting possibilities, it also raises several ethical concerns that must be addressed. The potential for privacy invasion and misuse of personality data is a significant risk. Therefore, transparent data handling procedures, strict compliance with data protection regulations, and the continuous updating and validation of models are essential to prevent bias and ensure fairness. Ensuring that personality profiling does not unfairly target or discriminate against individuals or groups is crucial for the ethical implementation of these technologies.

In conclusion, this study has highlighted the utility of deep learning models in uncovering the psychological underpinnings of cybersecurity risks. The integration of psychological profiling with cybersecurity not only enhances threat detection but also offers a novel preventative approach. By addressing the inherent psychological drivers of malicious online behavior, this research bridges a crucial gap between human psychology and technological defenses. However, the success of these approaches hinges on careful consideration of ethical implications and rigorous empirical validation to ensure that the models are fair, effective, and widely applicable in real-world scenarios. As deep learning continues to evolve, its potential to transform cybersecurity through the lens of human psychology offers a promising avenue for future research and application.

REFERENCE LIST

Abood, N., (2019) Big five traits: A critical review. *Gadjah Mada International Journal of Business*, 21(2), 159-186.

Alassiri, R., Abukhodair, F., Kalkatawi, M., Khashoggi, K. & Alotaibi, R., (2022) COVID-19 diagnosis from chest CT scan images using deep learning. *Revista Română de Informatică și Automatică*, 32(3), 65-72. doi: 10.33436/v32i3y202205

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

Cristescu, I., Ciupercă, E.M. & Cîrnu, C.E., (2022) Rolul trăsăturilor de personalitate în atacurile de ingineriesocială. *Revista Română de Informatică și Automatică*, 32(1), 113-122. doi:10.33436/v32i1y202209

Feng, J., Zhang, R., Chen, D., Shi, L. & Xiao, C., (2023) Label-based topic modeling to enhance medical triage for medical triage robots. *Studies in Informatics and Control*, 32(4), 37-48. doi: 10.24846/v32i4y202304

Floroiu, I. (2024) Big5PersonalityEssays: Introducing a Novel Synthetic Generated Dataset Consisting of Short State-of-Consciousness Essays Annotated Based on the Five Factor Model of Personality. *arXiv.org*. https://arxiv.org/abs/2407.17586

Floroiu, I. & Timisică, D. (2024) O analiză heideggeriană a modelelor bazate pe transformeri generativi preantrenați. *Revista Română de Informatică și Automatică*, 34(1), 13–22. doi:10.33436/v34i1y202402

Floroiu, I., Timisică, D., Radu, M. and Boncea (2023) Automated diagnosis of breast cancer using deep learning. *Romanian Journal of Information Technology and Automatic Control*, 33(3), 99–112. doi: 10.33436/v33i3y202308

Floroiu, I., Floroiu, M., Niga, A.-C. & Timisică, D. (2024) Remote Access Trojans Detection Using Convolutional and Transformer-based Deep Learning Techniques. *Romanian Cyber Security Journal*, 6(1), 47–58. doi:https://doi.org/10.54851/v6i1y202405

Gavrilă, V., Băjenaru, L., Dobre, C. & Tomescu, M., (2021) Towards the development of a Romanian lexicon for the analysis of emotions in the literary works of canonical authors. *Studies in Informatics and Control*, 30(2), 111-120. doi: 10.24846/v30i2y202110

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D. & Graepel, T., (2014) Manifestations of user personality in website choice and behaviour on online social networks. *Machine learning*, 95, 357-380.

Kosinski, M., Stillwell, D. & Graepel, T., (2013) Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15), pp. 5802-5805.

Majumdar, M. (2023) essays_csv dataset. www.kaggle.com/datasets/manjarinandimajumdar/essayscsv [Accessed 22nd July 2023]

Mikolov, T., Chen, K., Corrado, G. & Dean, J., (2013) Efficient estimation of word representations in vector space. *arXiv:1301.3781*. https://arxiv.org/abs/1301.3781

Ones, D.S. & Viswesvaran, C., (2003) The big-5 personality and counterproductive behaviors. Misbehaviour and Dysfunctional Attitudes in Organizations. *Palgrave Macmillan*, 211-249.

Paraschiv, E.-A. & Cîrnu, C.-E., (2024) Between the lines: generating, detecting and defending against textual deepfakes. *Romanian Cyber Security Journal*, 6(1), 3-13. doi: 10.54851/v6i1y202401

Paraschiv, E.-A. & Sultana, A.-E., (2024) Utilizarea potențialului Vision Transformers pentru clasificarea îmbunătățită a imaginilor OCT. *Revista Română de Informatică și Automatică*, 34(2), 97-111. doi: 10.33436/v34i2y202408

Rogers, M., Smoak, N.D. & Liu, J., (2006) Self-reported deviant computer behavior: A big-5, moral choice, and manipulative exploitive behavior analysis. *Deviant behavior*, 27(3), 245-268.

Silvia, P.J. & Christensen, A.P., (2020) Looking up at the curious personality: Individual differences in curiosity and openness to experience. *Current Opinion in Behavioral Sciences*, 35, 1-6.

Sudzina, F. & Pavlicek, A., (2020) Virtual offenses: Role of demographic factors and personality traits. *Information*, 11(4), p. 188.

Van de Weijer, S.G. & Leukfeldt, E.R., (2017) Big five personality traits of cybercrime victims. *Cyberpsychology, Behavior, and Social Networking*, 20(7), 407-412.

Wiggins, J.S. ed., 1996. The five-factor model of personality: Theoretical perspectives. *Guilford Press*.

Wilmot, M.P. & Ones, D.S., 2019. A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences*, 116(46), pp. 23004-23010.

Yin, W., Kann, K., Yu, M. & Schütze, H., (2017) Comparative study of CNN and RNN for natural language processing. *arXiv:1702.01923*.