DeepQSP: Identification of Quorum Sensing Peptides Through Neural Network Model

Submitted by

Md. Ashikur Rahman

ID: 242-44-013

Department of Software Engineering

Daffodil International University

Supervised by

Dr. Imran Mahmud

Professor and Head

Department of Software Engineering

Daffodil International University

This Thesis paper has been submitted in fulfillment of the requirements for the degree of Masters of Science in Software Engineering.

Summer 2025

APPROVAL

This thesis titled on "DeepQSP: Identification of Quorum Sensing Peptides Through Neural Network Model", submitted by Md. Ashikur Rahman, ID: 242-44-013 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Masters of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

---------- Chairman

DR. A. H. M. SAIFULLAH SADI
Professor
Department of Software Engineering
Daffodil International University

---------- Internal Examiner 1

DR. RUBAIYAT ISLAM
Associate Professor
Department of Software Engineering
Daffodil International University

---------- Internal Examiner 2

Afsana Begum
Assistant Professor
Department of Software Engineering
Daffodil International University

---------- External Examiner

ISHTIAK HUSSAIN
Project Manager
Dynamic Solution Innovators (DSi)

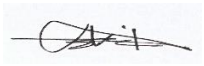
Declaration

I hereby declare that this work was carried out under the guidance of Dr. Imran Mahmud, Professor and Head of the Department of Software Engineering, Daffodil International University. I also confirm that this work has not previously been submitted, either in whole or in part, for the awarding of any Master's degree or other qualification.

Supervised By

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Daffodil International University

Submitted by



Md. Ashikur Rahman
ID: 242-44-013
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

The motivation for this study, DeepQSP: Identification of Quorum Sensing Peptides Through a Neural Network Model, was simply my curiosity and desire for gaining knowledge and learning more about this area. First of all, I would like to express my sincere appreciation to the God Almighty, who is All-Knowing, for supporting me with wisdom, strength, and perseverance through this process. Without His blessing I would not be able to have accomplished this study. I am very thankful to my amazing parents. Without their love, support, and encouragement, I would not have been able to complete this study. The love, support, and encouragement of others is what I have become as a result of all of the people in my life. I would like to thank to Dr. Imran Mahmud, Professor and Head of the Department of Software Engineering, Daffodil International University. His supervision, inspiration, and guidance was critical to this research. I would also like to express my appreciation and thanks to all my beloved teachers, especially those that I have mentioned in the presentation. Their commitment and everything they have taught both my classmates and I have been tremendously important in my academic journey and career. I would like to thank Daffodil International University for providing the resources and facilities, and academic setting to put together the work. I would like to acknowledge my classmates and fellows of Daffodil International University for their encouragement, cooperation and support during this process.

ABSTRACT

Objective: QSPs (quorum sensing peptides) are short signaling molecules that play a crucial role in microbial signaling. They allow bacteria to work together as a group, for actions such as in the establishing of biofilms and modulating virulence. When we can study the QSPs we can come to understand more of these biological processes. There are already clinical and laboratory procedures used to detect QSPs but they tend to be costly and lengthy to perform. Methods: The research advances a novel approach, known as DeepQSP, for detecting QSP. The approach consists of two extraction methods: Latent Semantic Analysis (LSA), a word feature extraction approach, and the classical amino-based extraction method of Pseudo Amino Acid Composition (PAAC), in addition to a convolutional neural network (CNN) classifier. Results: The DeepQSP model's performance was examined using a 440 peptide sequence dataset with the following results: accuracy 0.9697, sensitivity 0.9655, specificity 0.9730, and a Matthews correlation coefficient (MCC) score of 0.9385. The use of LSA and PAAC enhances peptide sequence representation. A CNN extracts and learns complex patterns effectively for accurate QSP recognition. Discussion: These findings demonstrate the efficiency of the DeepQSP method, which presents an important resource for investigating microbial interactions and quorum sensing. The identification of QSPs is crucial for microbiology and bioengineering, since understanding the interactions between microorganisms has a direct impact on understanding their potential as communicators.

Keywords — Quorum sensing peptide, Convolutional neural network, Latent semantic analysis, Microorganisms, Bioengineering.

Table of Contents

Table of Figures

Table of Tables

CHAPTER 1: INTRODUCTION

Quorum sensing (QS) is a crucial biological process used by microorganisms where bacterial cells communicate and regulate genes by the transfer of particular chemical signals (Fuqua et al. 1994, MB 2001). Quorum Sensing Peptides (QSP) are small signaling molecules produced and released by bacteria to facilitate communication among microbial community members (Dunny et al. 1997, Waters et al. 2005). These peptides are important for regulating different bacterial behaviors, such as biofilm formation, virulence factor production, and gene expression (Rutherford et al. 2012). Quorum sensing enables bacteria to coordinate actions based on the number of their peers by sensing the concentration of the QSP within their local environment. When the threshold is reached, a reaction is triggered in the bacterial assemblage for the bacteria to act as one unit. Quorum sensing is an interesting process that enables bacteria to adapt to environmental changes and respond to changes in conditions as a group. Researchers are exploring QSP for potential application in fields like medicine, biotechnology, food production, agriculture, and environmental science (Mangwani et al. 2012, Wu et al. 2020, Horswill et al. 2007).

QSP can support crop health with environmentally friendly methods of managing plant disease through the inhibition of quorum sensing in pathogens. In particular in the food production industry, these QSP activities can inhibit bacterial action and can also slow the unwanted formation of biofilms that can inhibit safety and quality of food products (Abbamondi and Tommonaro 2022, Skandamis and Nychas 2012). Researchers are also looking into the use of QSP to inform new antibacterial treatments; as these peptides can inhibit cell-to-cell communication and help reduce pathogenicity of some bacteria, without elimination. This method has potential to lessen problems associated with the wide-spread use of antibiotics and antibiotic resistance, which are increasingly problematic in modern medicine (Bramhachari 2019). Furthermore, QSP has applications in bioprocessing and bioremediation, to improve the efficacy of microbial fermentation processes and control the activity of genetically modified bacteria, for either the biosynthesis of target products or degradation of environmental pollutants (Sarkar et al. 2020). QSP is important across a range of scientific disciplines, from food and health to agriculture and environmental science. Therefore, a clear and cost-effective means of eliciting and measuring QSP is paramount to research and the health of humanity. Traditional

laboratory methodology for QSP or medicinal peptide sequences is costly and time consuming. In this case, the utilization of machine learning (ML) and computational biology will be paramount in resolving this complication. ML and advanced computational modeling can provide more accurate QSP and protein or peptide sequences prediction (Alsanea et al. 2022, Pu et al. 2021).

A number of studies have focused on the predictions of Quantitative Structure-Activity Relationship (QSARs) classification. However, the imitation of QSP compounded from G3- and K-models must be further developed. While models have been effective in the development of predictions, these models rely on the traditional and antiquated methods of feature extraction from amino acid variants. This will now include newly developed models that use word embedding methods of feature extraction which will show more subtle differences in the predictions. Above and beyond whatever amount of predictions has occurred, exploring models other types of machine learning and particularly deep learning methods often yield stronger and more accurate models. Given the wide application of QSP studies in a historically important fields, it is highly important there is more reliable, cost effective predictive modeling for QSP. Studies are underway to provide a QSP identification to the scientific community that will aim for high accuracy and effectiveness, surpassing the previous performance of the QSP models.

The research contribution aims to enhance QSP prediction performance. This research combines several word embedding feature extraction methods with some amino acid feature extractors to improve prediction resilience. The machine learning and deep learning algorithms used to build prediction models further strengthen this model and result in improved accuracy and power of the QSP prediction. The major novelty is introducing word embedding techniques in conjunction with traditional amino acid feature extraction methods; not many studies have explored this for QSP prediction. This new approach will capitalize on both, providing more relevant and accurate QSP prediction models, and machine learning expands the current capabilities of predicting QSP.

CHAPTER 2: LITERATURE REVIEW

Over the past few years, researchers have conducted numerous studies to make predictions about QSP using ML and computational models. Rajput and Kumar 2015 published the first computational model for predicting the QSP, called QSPpred. The authors encoded the sequence into different variables including amino acid composition (AAC), amino acid residue position (ACRP), sequence motifs, physicochemical properties (PCP), and the Grand average of hydropathy (GRAVY). The authors used a support vector machine (SVM) as the classifier model with remarkably high accuracy (93.00% and Mathew's correlation coefficient (MCC) of 0.8600 (Rajput and Kumar 2015). Although the SVM produced higher accuracy scores, its reliance on several different encodings made it computationally expensive. The SVM classifier worked well, but would likely struggle to represent complex patterns in the data in the same way that more advanced machine learning techniques like deep learning would. Wei et al. (2020) developed a QSP model, utilizing Random Forest (RF), and research feature representation learning known as QSPred-FL. QSPred-FL produced an accuracy of 94.30% and MCC of 0.8850 (Wei et al. 2020). Although QSPred-FL demonstrated a high level of accuracy, RF models struggle at a high-dimensional space and are prone to generalizing poorly with unseen data. Though the feature representation learning methodology behind this model is advantageous, it still resorts to conventional ligeding techniques that could not provide us with a complete account of QSP sequence complexity. In another study, Charoenkwan et al. (2019) introduced yet another SVM classifier based model iQSP predictor which utilizes a PCP-based encoding technique, with maximum accuracy of 93.00% and MCC of 0.8600 (Charoenkwan et al. 2019). Importantly, while PCP-based encoding may represent an advance, the iQSP predictor plateaued pol performance when viewed as a continuation of earlier resaerch. The implementation of SVM also limits more complex learning in space of QSP sequences and because this, in tandem with several encoding techniques does not capture all relevant features. In 2022 Sivaramakrishnan and Ponraj introduced EnsembleQS a stacking based ensemble learning model for detecting QSP. The features implements Dipeptide Composition (DPC), Dipeptide Deviation from Expected Mean (DEM), AAC, and Tripeptide Composition (TPC). The EnsembleQS predictor does not disappoint, providing an accuracy of 93.40% and a MCC of 0.9100

(Sivaramakrishnan and Ponraj 2022). Ensemble-based modelling, internal or external improves prediction performance but incurs an increase in computational complexity, which has a risk of overfitting. In addition, although the feature encoding methods are varied, they still remain traditional approaches and may not account for the more in-depth contextual information from sequences. Charoenkwan et al. (2023) proposed the PSRQSP model for QSP prediction which took into account propensity scores for the 20 amino acids and for 400 dipeptides. In this work, propensity scores were implemented with a scoring card method to construct PSRQSP. The PSRQSP model appears to have achieved the highest accuracy previously reported for amino acid sequences at 94.44% accuracy using the same dataset (Charoenkwan et al. 2023). Although this is an innovative approach and a successful study, the model relies on propensity scores and scoring card methods, which may limit its generalizability to other datasets. Furthermore, improvements in accuracy do not promote innovation in feature extraction or machine learning methods.

CHAPTER 3: METHODS AND MATERIALS

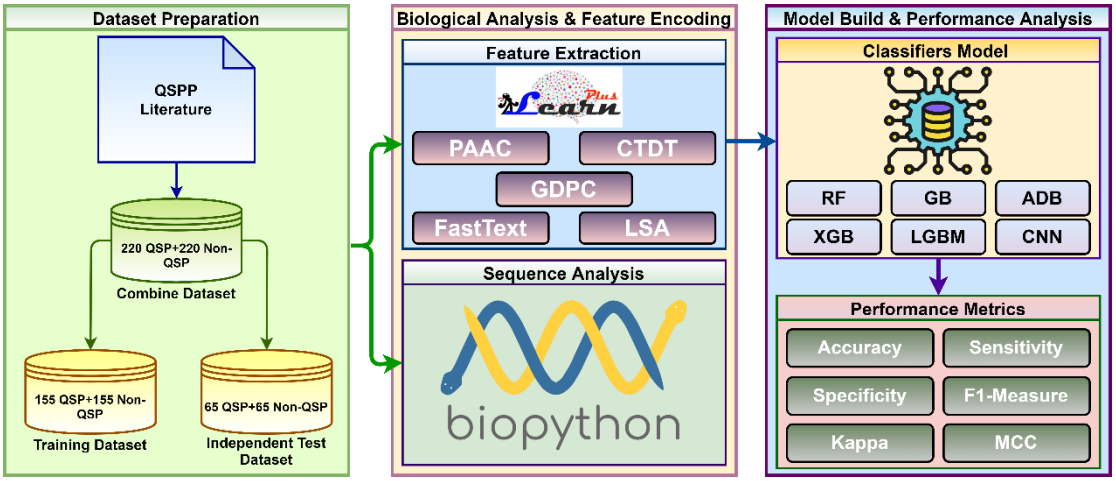


Figure 3.1: Research methodology of the study and structural architecture for developing the proposed model, which has been named after DeepQSP.

3.1 Dataset Description

In order to build a more robust and powerful QSP predictor in this study, we used previously published datasets developed by other researchers (Rajput and Kumar 2015, Charoenkwan et al. 2019, Wei et al. 2020, Sivaramakrishnan and Ponraj 2022, Charoenkwan et al. 2023). The datasets have an overall total of 440 peptide sequences. Of the total, 220 were QSP sequences and 220 were non-QSP sequences. We divided the whole dataset into training and independent testing datasets for this study. The training dataset consisted of 310 sequences: 155 QSP sequences and the rest non-QSP sequences. The independent test dataset was made of 130 sequences, again equally distributed between QSP and non-QSP sequences.

3.2 K-fold Cross Validation (CV)

Cross-Validation (CV) with 5-Folds is a common way to assess performance in machine learning. With this approach, you take a dataset and split it into five (more or less) equal parts, or folds. You then train and test the model five times (what is known as five-fold cross-validation), using four folds for training and one fold for testing. This works particularly well for classification problems, because it helps reduce overfitting and gives a more stable estimate of the performance of the model.

3.3 Feature Encoding

3.3.1 Pseudo-Amino Acid Composition

Pseudo amino acid composition (PAAC) is a method using computers to represent

protein or peptide sequences. It incorporates various properties of amino acids to assist in bioinformatics tasks such as predicting structures and functions of proteins (Chou 2009, Chou 2005, Chou 2001). The PAAC can be defined as:

$$X_c = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{\lambda}\theta_j}, (1 < c < 20)$$

$$X_c = \frac{w\theta_{c-20}}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{\lambda}\theta_j}, (21 < c < 20 + \lambda)$$

The weighting factor for the sequence-order effect, denoted as w is typically set to a value of 0.05 in iLearnPlus, as recommended by Chou et al. (2001)..

3.3.2 Composition, Transition, Distribution

The CTDT (Composition, Transition, Distribution, Transition) approach is used for biologically relevant feature extraction in bioinformatics. Here, the authors propose four different feature types to describe a protein or peptide sequence. This is helpful for a range of computational analyses (Mostafa et al., 2022; Wang et al., 2020). CTDT works as follows:

$$T(r,s) = \frac{N(r,s) + N(s,r)}{N-1},$$

$$r,s \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\}$$

$N(r,s)$ and $N(s,r)$ represent the counts of dipeptides encoded as "*rs*" and "*sr*", respectively, in the sequence. "*N*" corresponds to the length of the sequence.

3.3.3 Grouped Dipeptide Composition

An alternative version of the DPC descriptor, called the Grouped Dipeptide Composition (GDPC) encoding, includes a total of 25 dimensions (Chen et al. 2020, Chen et al. 2018). GDPC is defined as:

$$f(r,s) = \frac{N_{rs}}{N-1}, \qquad r,s \in \{g1, g2, gg3, 4, g5\}$$

Here, N_{rs} denotes the quantity of tripeptides produced by the amino acid type groups denoted by the letters "*r*" and "*s*", and *N* is the total length of a peptide or protein sequence.

3.3.4 FastText

FastText is an algorithm for word embedding that maps words to continuous vectors. This method allows for capturing the meanings of words as well as the relationships between words. FastText is known for being fast and efficient in dealing with out-of-vocabulary words, making it useful for various natural language processing

challenges (Didi et al. 2022, Selva Birunda and Kanniga Devi 2021). In the last few years, FastText has grown popular with researchers for encoding biological sequences (Le and Huynh 2019, Shi and Chen 2021, Lv et al. 2021).

The FastText works based on the following equation:

$$E(w) = \sum_{g \epsilon G_w} z_g$$

Here, $E(w)$ represents the word vector for word w, G_w is the set of subword (character) n-grams for the word w, z_g represents the vector for subword g in G_w.

3.3.5 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a commonly used word embedding technique. It represents words' meanings in a high-dimension space - via vector space - which models word relationships and meanings based on patterns of co-occurrence in a large text corpus (Naili et al. 2017, Joshi et al. 2016). LSA word embedding techniques have been investigated in studies of biological sequence data encoding (Patrick et al. 2019, Liu et al. 2022). LSA has two components; a term-document matrix and a singular value decomposition (SVD) (Tékouabou et al. 2022). These two terms are defined as follows:

$$M = \left[m_{11,} m_{12,} \dots, m_{1n,} \quad m_{21,} m_{22,} \dots, m_{2n,} \dots \dots \dots \dots \; m_{m1,} m_{m2,} \dots, m_{mn,}\right]$$

$$M = U * \Sigma * V\text{^}T$$

In this case, the matrices U, Σ, and $V\text{^}T$ have dimensions (m x k), (k x k), and (k x n), respectively. The chosen reduced dimensionality, 'k,' corresponds to the number of latent semantic dimensions selected for embedding.

3.4 Development of DeepQSP using Deep Learning

Deep learning is a specialized area of machine learning, and represents a highly effective approach in computational biology when analyzing complex biological datasets. Deep learning involves training artificial neural networks on a sufficient amount of data so that it can recognize patterns and make predictions. Here, deep learning methods are used to identify Quorum Sensing Peptides (QSPs), which are important in bacterial signaling. The approach leverages the ability of neural networks to process and learn from the intricate patterns in peptide sequences, enabling more accurate and efficient identification compared to traditional methods. The developed deep learning-based method is named DeepQSP, where the CNN layer is employed. Then the performances of the DeepQSP are compared with different machine learning

to ensure that the proposed DeepQSP is highly capable of identifying QSPs from protein sequences. The structural architecture of this study has been represented in Figure 3.4.1. CNN was used to develop DeepQSP for the identification of Quorum Sensing Peptides. The process starts with the input layer, which applies convolution using 64 filters of kernel size 2. This operation can be mathematically represented as (Chen et al. 2021, Darwish and Hassanien 2020):

$$y = f(W \times x + b)$$

where x is the input, W represents the weights of the filters, and b is the bias, x and denotes the convolution operation. The ReLU activation function is then applied for non-linear transformation. The ReLU activation function can be expressed as follows (Darwish and Hassanien 2020):

$$f(x) = \max(0, x)$$

Subsequent hidden layers, with 128 filters each and ReLU activation, further process these features. The pooling layer, with a kernel size of 4, reduces the spatial dimensions (downsampling), which can be represented as (Darwish and Hassanien 2020, Lin et al. 2018):

$$P(x)_{ij} = x_{4i+k,4j+l}$$

Where $P(x)_{ij}$ is the pooled output. Afterward, a flattened layer is employed to convert all the previous layer's output from pooled feature maps into a single long continuous linear vector. The network then includes dense layers with 128 and 64 nodes respectively, again using ReLU activation, to interpret these features. The final output layer with a sigmoid activation function, is used for binary classification. The mathematical expression of sigmoid function is as follows (Darwish and Hassanien 2020, Lin et al. 2018, Sandhiya and Palani 2020):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The model is trained over 40 epochs with batches of 64 samples each, optimizing the network's ability to identify QSPs from the input data accurately. This architecture is a strategic blend of convolutional layers and dense layers, using non-linear activation functions for complex pattern recognition in peptide sequences. The structural architecture of the proposed model is illustrated in Figure 3.4.1. The parameters used in this study are presented in Table 3.4.1.

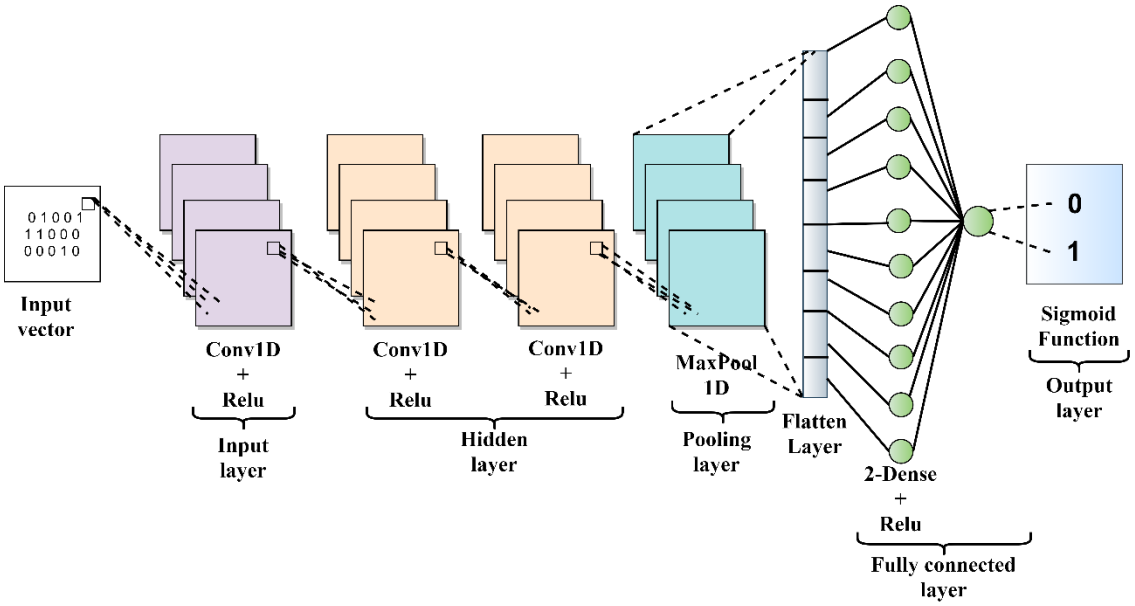


Figure 3.4.1: Structural architecture of the proposed model DeepQSP, which has been developed using CNN architecture.

Table 3.4.1: Parameters of the applied models of this study.

Model	Parameter			
CNN	Layer	Filters	Kernel Size/Pool size	Activation function
	Input	64	2	relu
	Hidden-1	128	2	relu
	Hidden-2	128	2	relu
	Pooling		4	
	Dense-1	128		relu
	Dense-2	64		relu
	Output	1		sigmoid
	epoch: 40; batch size: 64			

3.5 Evaluation Metrics

Assessing the performance of deep learning (DL) and machine learning (ML) models is necessary to evaluate their utility and predictive ability. The choice of evaluation metrics, however, depends on the specific task (classification, regression, etc.). In this research, six evaluation metrics were used to analyze the performance of the implemented models. Accuracy is defined as the ratio of a correct predicted instances

to the total number of instances (Charoenkwan et al. 2022). Specificity can be defined as the ratio of correctly predicted negative instances to the total sum of actual negative instances (Ali et al. 2021). Sensitivity is the number of actual positive instances that the algorithm actually predicted correctly as positive (i.e., true positives), divided by the total actual positive instances. (Ali et al. 2021).

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

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

Another performance metric is F1-Measure. The F1-Measure happens to be a statistically better evaluated model where the integration of precision and recall will make it powerful and effective in classification tasks (Ali et al. 2021). Kappa Statistics assesses agreement between raters or classifiers, accounting for chance agreement, in various tasks such as inter-rater reliability or classification model evaluation (Mohamed 2017). The range of Matthews Correlation Coefficient (MCC) is -1 to +1, effectively quantifying the strength of association between two variables (Ali et al. 2021).

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Kappa\ Stat = \frac{observed\ accuracy - \ expected\ accuracy}{1 - expected\ accuracy}$$

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

Here, True positives are denoted by *TP*, whereas true negatives are denoted by *TN*, false positives by *FP*, and false negatives by *FN*.

CHAPTER 4: RESULT ANALYSIS

In the research, a two-step analytical strategy was used. First, the biological analysis of peptide sequences was completed with the help of the open-source Biopython software (biopython.org). Then, version 3.10.12 of Python was used in the Google Colab development environment to build and analyze performance of the proposed model.

4.1 Analysis of Peptide Sequence

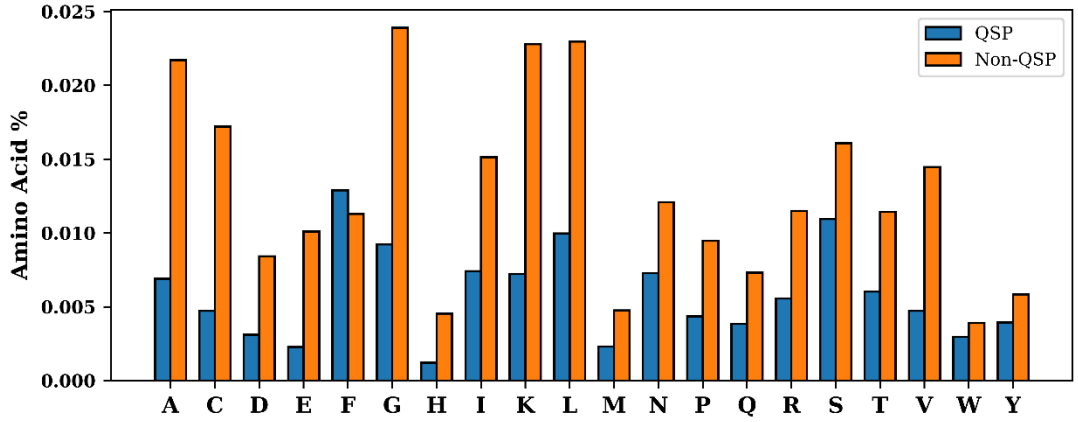


Figure 4.1: Percentage of the amino acids in the used QSP dataset.

Figure 4.1 shows the amino acid percentage of our used datasets. Based on the figure, the F (phenylalanine) has the highest percentage value among the twenty amino acids, considering the QSP sequence. G (glycine) shows the maximum percentage among other amino acids for the non-QSP sequence. The lowest percentage value among the amino acids is W (tryptophan) for the non-QSP. And for the QSP sequence, the H (histidine) has the lowest percentage. Here it has been seen that the percentage value of the non-QSP sequence is significantly higher than the QSP sequence.

4.2 Analysis of the Classifiers Algorithm Result

In this study, 5-fold CV was applied to construct classification models for the training dataset. The utilized datasets are balanced, with an equal number of QSP and non-QSP sequences, eliminating the need for any balancing method. After applying 5-fold CV to the training dataset, the classification model was evaluated on an independent dataset. In addition, the feature extractor was merged, combining 3 amino acid property-based extractors with 2-word embedding techniques, resulting in six

different combinations of feature extraction methods. The results of the 5-fold CV and independent test are represented in Table 4.1 and Table 4.2.

Table 4.1: 5-fold CV results of the different combined feature extractor.

Extractor	Classifiers	Accuracy	Sensitivity	Specificity	F1-Measure	MCC	Kappa
LSA+ PAAC	RF	0.7857	0.8366	0.7355	0.7950	0.5748	0.5717
	XGB	0.8084	0.8104	0.8064	0.8078	0.6169	0.6169
	LGBM	0.8084	0.8104	0.8064	0.8078	0.6169	0.6169
	GB	0.8344	0.8301	0.8387	0.8328	0.6688	0.6688
	ADB	0.8279	0.8758	0.7806	0.8349	0.6592	0.6560
	CNN	0.9672	0.9667	0.9677	0.9667	0.9344	0.9344
LSA+ CTDT	RF	0.8006	0.7671	0.8312	0.7860	0.6003	0.5996
	XGB	0.7879	0.7397	0.8312	0.7687	0.5744	0.5729
	LGBM	0.8072	0.8219	0.7937	0.8027	0.6150	0.6144
	GB	0.8235	0.8082	0.8375	0.8138	0.6462	0.6461
	ADB	0.7908	0.7603	0.8187	0.7762	0.5805	0.5801
	CNN	0.9180	0.9630	0.8823	0.9123	0.8357	0.8398
LSA+ GDPC	RF	0.8377	0.8456	0.8302	0.8344	0.6755	0.6752
	XGB	0.8961	0.8725	0.9182	0.8904	0.7923	0.7917
	LGBM	0.9091	0.8926	0.9245	0.9048	0.8181	0.8178
	GB	0.8896	0.8658	0.9119	0.8836	0.779	0.7787
	ADB	0.8474	0.8121	0.8805	0.8374	0.6951	0.6939
	CNN	0.9672	1.0000	0.9355	0.9677	0.9365	0.9345
FastText+ PAAC	RF	0.8562	0.8973	0.8187	0.8562	0.7160	0.7130
	XGB	0.8693	0.8699	0.8687	0.8639	0.7382	0.7382
	LGBM	0.8529	0.8630	0.8437	0.8485	0.7061	0.7057
	GB	0.8660	0.8767	0.8562	0.8619	0.7322	0.7319
	ADB	0.8431	0.8699	0.8187	0.8410	0.6880	0.6865
	CNN	0.9344	0.9310	0.9375	0.9310	0.8685	0.8685
FastText+ CTDT	RF	0.8006	0.7671	0.8312	0.7860	0.6003	0.5996
	XGB	0.7876	0.7397	0.8312	0.7687	0.5744	0.5729
	LGBM	0.8072	0.8219	0.7937	0.8028	0.6150	0.6144

	GB	0.8235	0.8082	0.8375	0.8138	0.6462	0.6461
	ADB	0.7908	0.7603	0.8187	0.7762	0.5805	0.5801
	CNN	0.9508	0.9687	0.9310	0.9538	0.9017	0.9012
FastText+ GDPC	RF	0.8333	0.8278	0.8387	0.8306	0.6666	0.6666
	XGB	0.8431	0.8410	0.8452	0.8410	0.6862	0.6862
	LGBM	0.8268	0.8212	0.8322	0.8239	0.6535	0.6535
	GB	0.8497	0.8344	0.8645	0.8456	0.6994	0.6992
	ADB	0.8203	0.8079	0.8322	0.8160	0.6405	0.6404
	CNN	0.9016	0.9286	0.8788	0.8965	0.8047	0.8030

The CV result of the classifier models is shown in Table 4.1. The information provided in the table shows that the LSA+GDPC feature extractor method achieves the highest result considering all the performance metrics, except for only the specificity of the CNN classification model. The accuracy is 96.72%, the sensitivity is 1.00, the specificity is 0.9355, the F1-measure is 0.9677, and the MCC and Kappa are 0.9365 and 0.9345, respectively, for the LSA+GDPC encoding method on CNN classifier. The LSA+PAAC feature encoding method shows the maximum specificity score. The accuracy is 96.72%, which is equal to the LSA+GDPC encoding method's accuracy, but the sensitivity is 0.9667, the specificity is 0.9677, the F1-measure is 0.9667, and the MCC and Kappa are 0.9344 for the LSA+PAAC encoding method. The lowest performance has been shown by the RF classifier on the LSA+PAAC extractor technique. The minimum accuracy, sensitivity, and specificity is 0.7857, 0.8366, and 0.7355 respectively. 0.7950, 0.5748, and 0.5717 are the least F1-Measure, MCC, and Kappa scores.

Table 4.2*:* Independent Test results of the different combined feature extractors.

Extractor	Classifiers	Accuracy	Sensitivity	Specificity	F1-Measure	MCC	Kappa
LSA+ PAAC	RF	0.8485	0.8955	0.8000	0.8571	0.6994	0.6965
	XGB	0.8561	0.8358	0.8769	0.8550	0.7130	0.7122
	LGBM	0.8712	0.8658	0.8769	0.8722	0.7425	0.7424
	GB	0.9015	0.9104	0.8923	0.9037	0.8030	0.8029
	ADB	0.9091	0.9254	0.8923	0.9118	0.8184	0.8180
	CNN	0.9697	0.9730	0.9655	0.9730	0.9385	0.9385

LSA+ CTDT	RF	0.8257	0.8167	0.8333	0.8099	0.6492	0.6491
	XGB	0.7954	0.8167	0.7778	0.7840	0.5920	0.5903
	LGBM	0.8257	0.8500	0.8055	0.8160	0.6529	0.6510
	GB	0.8182	0.8333	0.8055	0.8064	0.6365	0.6353
	ADB	0.8030	0.7833	0.8194	0.7833	0.6028	0.6028
	CNN	0.9318	0.9077	0.9552	0.9291	0.8644	0.8635
LSA+ GDPC	RF	0.8257	0.8871	0.7714	0.8271	0.6592	0.6531
	XGB	0.9015	0.9032	0.9000	0.8960	0.8026	0.8025
	LGBM	0.9470	0.9516	0.9428	0.9440	0.8937	0.8936
	GB	0.9242	0.9032	0.9428	0.9180	0.8480	0.8476
	ADB	0.9091	0.9193	0.9000	0.9048	0.8182	0.8178
	CNN	0.9621	0.9851	0.9385	0.9635	0.9251	0.9242
FastText+ PAAC	RF	0.8257	0.8630	0.7797	0.8456	0.6465	0.6458
	XGB	0.8257	0.8356	0.8135	0.8414	0.6482	0.6481
	LGBM	0.8182	0.8630	0.7627	0.8400	0.6310	0.6298
	GB	0.8485	0.8493	0.8474	0.8611	0.6948	0.6945
	ADB	0.8333	0.8630	0.7966	0.8513	0.6621	0.6618
	CNN	0.9697	0.9855	0.9524	0.9714	0.9396	0.9392
FastText+ CTDT	RF	0.8712	0.8356	0.9152	0.8777	0.7466	0.7424
	XGB	0.8712	0.8767	0.8644	0.8827	0.7400	0.7399
	LGBM	0.9015	0.8904	0.9152	0.9091	0.8026	0.8017
	GB	0.8864	0.8630	0.9152	0.8936	0.7742	0.7720
	ADB	0.8864	0.8082	0.9830	0.8872	0.7901	0.7749
	CNN	0.9318	0.9846	0.8806	0.9343	0.8687	0.8638
FastText+ GDPC	RF	0.8485	0.8529	0.8437	0.8529	0.6967	0.6967
	XGB	0.8409	0.8970	0.7812	0.8531	0.6844	0.6805
	LGBM	0.8788	0.9118	0.8437	0.8857	0.7583	0.7569
	GB	0.8864	0.9265	0.8437	0.8936	0.7742	0.7720
	ADB	0.8182	0.8676	0.7656	0.8310	0.6377	0.6350
	CNN	0.9394	0.9687	0.9118	0.9394	0.8805	0.8789

Table 4.2 shows the independent test result of this work. According to the data represented in the above table, that the LSA+PAAC and FastText+PAAC encoding methods achieve the maximum accuracy result, which is 96.97% on the CNN model. The sensitivity is 0.9730, the specificity is 0.9655, the F1-measure is 0.9730, and the Kappa and MCC are 0.9385 for the LSA+PAAC extractor on the CNN model. FastText+PAAC achieved the following results on the CNN classifier model:

sensitivity of 0.9855, specificity of 0.9524, F1-measure of 0.9714, Kappa of 0.9392, and MCC of 0.9396. The XGB classifier exhibited the lowest accuracy, recording 0.7954, with an MCC of 0.5920 and a Kappa score of 0.5903. On the contrary, the ADB classifier attained the lowest sensitivity and F1-Score, both registering at 0.7833, using the LSA+ CTDT feature extractor. The lowest specificity is 0.7714, derived by the RF model for LSA+ GDPC encoding method.

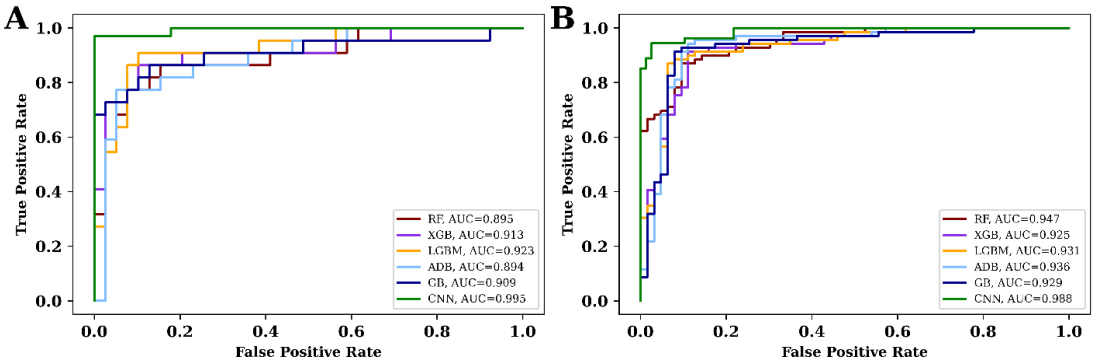


Figure 4.2.1: ROC curve for the different ML classifiers on LSA+PAAC feature encoding method. Subplot(A) shows the result of 5-flod CV and subplot(B) shows the result of the independent test.

The ROC curve with the AUC score of the six applied classifiers is represented in Figure 4.2.1. According to the figure, in subplot (A), the CNN achieved the maximum AUC score of 0.995 and the ADB achieved the lowest AUC of 0.893. On the other hand, for the independent test in subplot (B), the highest AUC score is 0.988, which is achieved by CNN, and the lowest AUC score is achieved by XGB, which is 0.925. In both subplots, the CNN model obtained the highest AUC score.

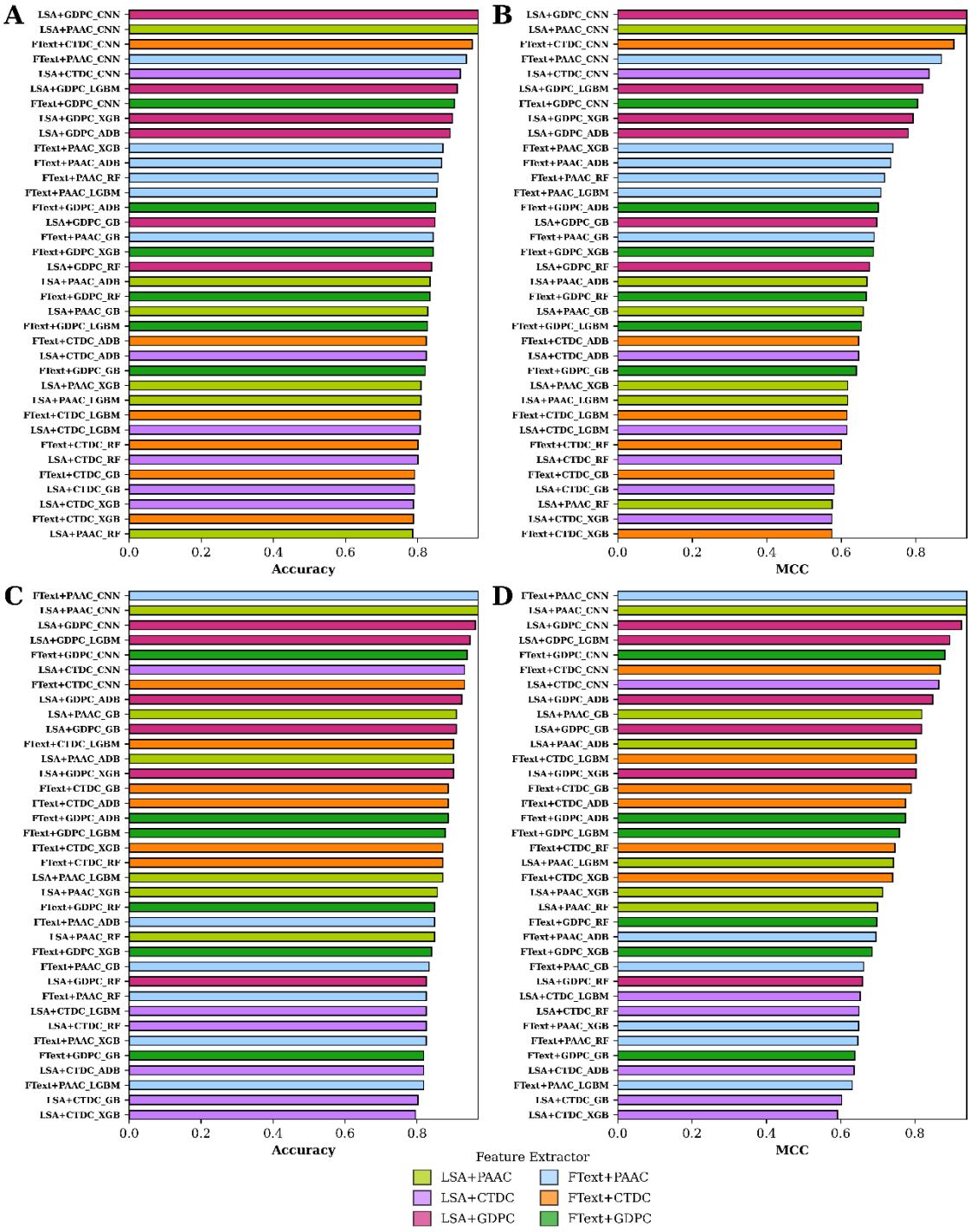


Figure 4.2.2: Comparison of the Accuracy and MCC of the combined feature encoding techniques and applied ML algorithm. Subplot(A) and subplot(B) for the 5-fold CV result. Subplot(C) and subplot(D) for the independent test result.

The comparison of the accuracy and MCC of the six feature extraction methods with the six applied classifier models (in total thirty-six models) are presented in Figure 4.3. Based on the information in the figure, the LSA+GDPC_CNN and LSA+PAAC_CNN models take the top two positions for a 5-fold CV, taking into

account the accuracy and MCC performance metrics. In the independent test, FastText+PAAC_CNN and LSA+PAAC_CNN are at the top of all the models for accuracy and MCC. LSA+PAAC_CNN are in both the 5-fold CV and independent tests. For the 5-fold CV accuracy and MCC, the LSA+PAAC_RF and FText+CTDC_XGB are in the bottom individually. Conversely, in both subplot (C) and subplot (D), LSA+CTDC_XGB is positioned at the last.

The sensitivity and specificity comparison of the thirty-six models are depicted in Figure 4.2.3. As indicated by the figure, for 5-fold CV specificity the LSA+PAAC_CNN model achieved the top among all the models. But in the sensitivity of the 5-fold CV, the LSA+PAAC_CNN is in the third position. Here, the LSA+GDPC_CNN model is at the top. In the independent test, FastText+CTDC_GB ranked highest for specificity, with LSA+PAAC_CNN coming in second place among all the models. For sensitivity in the independent test, FastText+PAAC_CNN took the lead, while LSA+PAAC_CNN secured the fourth position among all the utilized models.

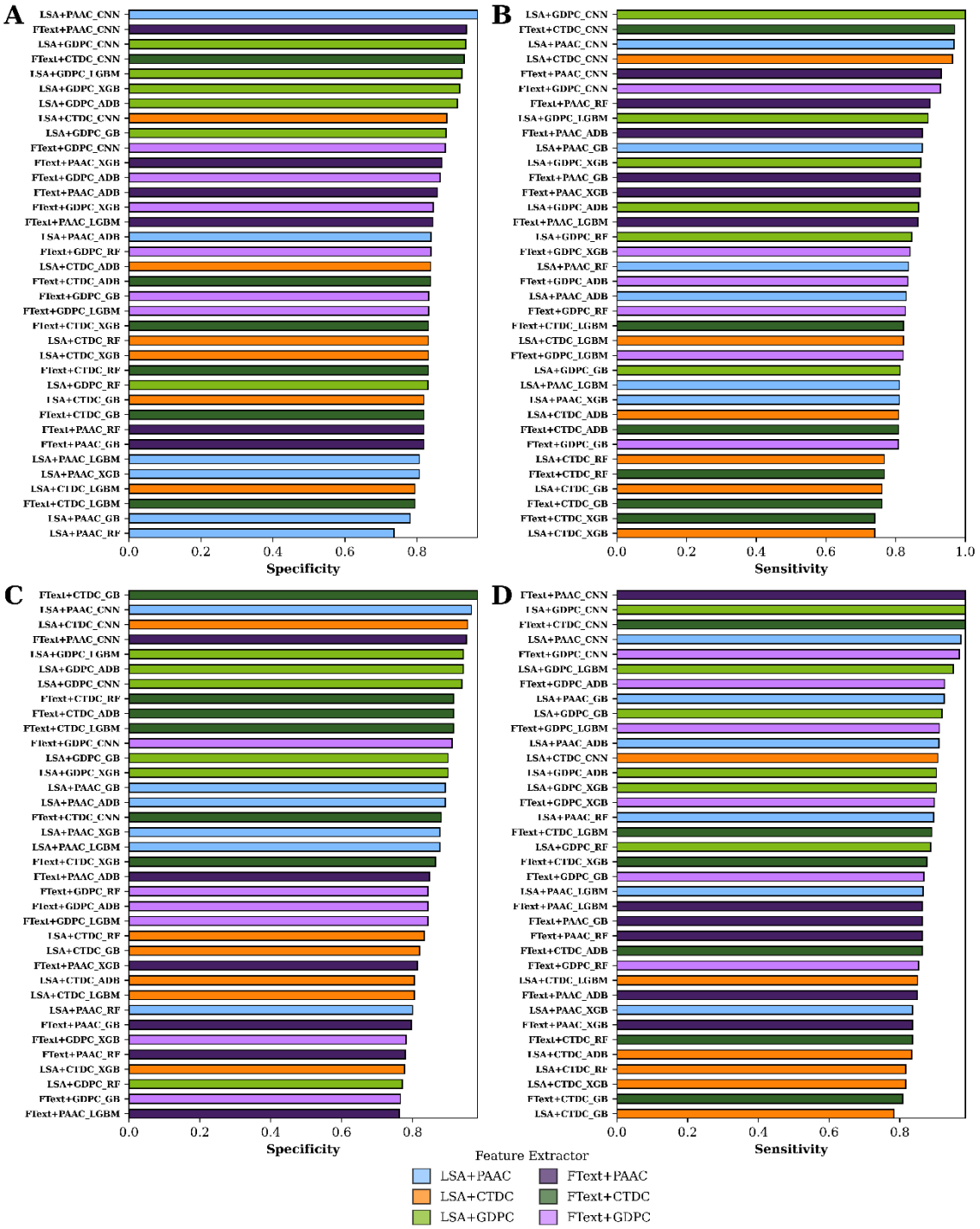


Figure 4.2.3: Compare the specificity and sensitivity of the combined feature encoding techniques and applied ML algorithm. Subplot(A) and subplot(B) for the 5-fold CV result. Subplot(C) and subplot(D) for the independent test result.

CHAPTER 5: DISCUSSION

The identification of quorum-sensing peptides is vital for comprehending and managing microbial activities and driving progress across diverse sectors, including healthcare, food, agriculture, drug discovery, biotechnology, and environmental science. Take account the analysis of the results discussed above, it can be concluded that the LSA+PAAC_CNN model demonstrates superior performance compared to all other models applied in the study. Though some models perform well for 5-fold CV and some models perform well in independent tests, the LSA+PAAC_CNN model outperforms on both the 5-fold CV and independent tests, considering all the evaluation metrics. The results of our developed DeepQSP prediction model on the independent test set are as follows: an accuracy of 0.9697, a specificity of 0.9730, a sensitivity of 0.9655, an F1-measure of 0.9730, and a kappa score and MCC score both equal to 0.9385. Additionally, the AUC score for our proposed DeepQSP model on the independent test is 0.988.

Table 5.1: Comparing DeepQSP with existing other QSP prediction models.

Prediction Model	Accuracy	Specificity	Sensitivity	Kappa	MCC
QSPpred (Rajput and Kumar 2015)	0.9000	-	-	-	0.8000
QSPred-FL (Wei et al. 2020)	0.9250	-	-	-	0.8600
iQSP (Charoenkwan et al. 2019)	0.9300	0.9350	0.9250	-	0.8600
EnsembleQS (Sivaramakrishnan and Ponraj 2022)	0.9340	-	-	-	0.9100
PSRQSP (Charoenkwan et al. 2023)	0.9444	1.0000	0.8820	-	0.8930
DeepQSP [this work]	**0.9697**	**0.9730**	**0.9655**	**0.9385**	**0.9385**

According to Table 5.1 and Figure 5.1, the performance of our proposed DeepQSP model significantly improves over the existing QSP prediction model. Our DeepQSP model outperforms other existing prediction models by a margin of 0.0253 to 0.0697 in terms of accuracy.

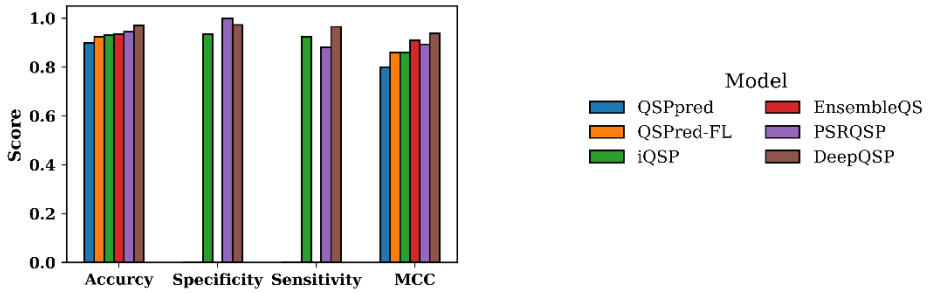


Figure 5.1: Performance comparison of DeepQSP with previous existing studies.

While our introduced DeepQSP outperforms other QSP predictor models, it still has limitations and future work. The limitations are: the sample size of the dataset used in this study is rather small. The dataset comprised only 440 dataset instances. A larger dataset would provide a better representation of QSP sequences that would help with better accuracy of the developed model when generalizing to unseen data. In addition, although the blend of word embedding and feature extraction from amino-acids seem to yield reasonably good results, there may be other feature extraction methods that could result in even better model performance. Although deep learning models similarly to CNNs are often seen as black boxes with high complexity, slight interpretation of the specific features or patterns underlying their predictions can be difficult. Our future work consists of developing a more trustworthy and robust QSP predictor model. We hope to collect a more robust, generalizable, and diverse QSP dataset; this would likely help in generalizing to performance across different microbial communities, or contexts. We are also planning to create a web-based server based on the developed DeepQSP predictor. This will serve as a user-friendly instrument for researchers and practitioners in bioengineering as well as the scientific community at large, as they will be able to utilize our model to develop new ways of implementing it. The next step will broaden-out incorporate more feature extraction methods, consisting of the structural and functional features of QSPs, with the goal of improving the predictive nature of the model. Additionally, we will broaden this work into transfer learning, hazing and utilize what we derive from relevant domains so that we can allow the model to generalize, and ultimately make better predictions with limited data. We will continue to focus our efforts on model optimization including the modeling design and training to reduce and enhance computational support and ease of use in real-life applications. Future work will also look into understandability of our DeepQSP model, so researchers can better understand which informing

features and characteristics indicate the most influence in predicting the model. Other methods, such as attention mechanisms or other explainable AI methods, can also be explored. Limitations of the DeepQSP model will be addressed and future work will continue to build on these contributions. These will all greatly contribute to the microbial communication literature and its applications in biotechnology, healthcare, and beyond.

CHAPTER 6: CONCLUSION

In conclusion, the combination of Conventional Neural Network models and word embedding feature encoding methodologies represent a novel approach for predicting Quorum Sensing Peptides. The new combination will help us study microbial communication better while moving forward to solve the problem of Quorum Sensing complexity. The model uses a Conventional Neural Network approach alongside word embedding techniques to perform research tasks effectively. The available application methods provided for Quorum Sensing Peptides open avenues for new healthcare therapies, reducing our reliance on broad-spectrum antibiotics in response to antibiotic resistance on a global scale. The study provides important insights about the subject which enables Quorum Sensing Peptides to spread across biotechnology, healthcare, and other industries. Our model achieved good performance, with an accuracy of 0.9697 and a Matthew's Correlation Coefficient of 0.9385. The research demonstrates a major breakthrough in microbial communication studies which benefits human health and environmental stability. Scientific development. Further research will focus on expanding the model to predict QSPs in more diverse microbial environments and exploring the integration of additional biological factors into the model. Additionally, future studies will aim to validate the model’s predictions experimentally, thereby strengthening its application in biotechnology, healthcare, and other fields.

CHAPTER 7: REFERENCES

Abbamondi, G.R. and Tommonaro, G., 2022. Research progress and hopeful strategies of application of quorum sensing in food, agriculture and nanomedicine. Microorganisms, 10(6), p.1192.

Ali, M.M., Ahmed, K., Bui, F.M., Paul, B.K., Ibrahim, S.M., Quinn, J.M. and Moni, M.A., 2021. Machine learning-based statistical analysis for early stage detection of cervical cancer. Computers in biology and medicine, 139, p.104985.

Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M. and Moni, M.A., 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, p.104672.

Alsanea, M., Dukyil, A.S., Afnan, Riaz, B., Alebeisat, F., Islam, M. and Habib, S., 2022. To Assist Oncologists: An Efficient Machine Learning-Based Approach for Anti-Cancer Peptides Classification. Sensors, 22(11), p.4005.

Bramhachari, P.V. ed., 2019. Implication of quorum sensing and biofilm formation in medicine, agriculture and food industry. Springer.

Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T. and Shoombuatong, W., 2019. iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides using informative physicochemical properties. International Journal of Molecular Sciences, 21(1), p.75.

Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Moni, M.A., Manavalan, B. and Shoombuatong, W., 2022. StackDPPIV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. Methods, 204, pp.189-198.

Charoenkwan, P., Chumnanpuen, P., Schaduangrat, N., Oh, C., Manavalan, B. and Shoombuatong, W., 2023. PSRQSP: An effective approach for the interpretable prediction of quorum sensing peptide using propensity score representation learning. Computers in Biology and Medicine, 158, p.106784.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.C. and Song, J., 2018. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics, 34(14), pp.2499-2502.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I. and Chou, K.C., 2020. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in bioinformatics, 21(3), pp.1047-1057.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S. and Miao, Y., 2021. Review of image classification algorithms based on convolutional neural networks. Remote Sensing, 13(22), p.4712.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics, 43(3), pp.246-255.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21(1), pp.10-19.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics, 6(4), pp.262-274.

Darwish, A., Ezzat, D. and Hassanien, A.E., 2020. An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. Swarm and evolutionary computation, 52, p.100616.

Didi, Y., Walha, A. and Wali, A., 2022. COVID-19 tweets classification based on a hybrid word embedding method. Big Data and Cognitive Computing, 6(2), p.58.

Dunny, G.M. and Leonard, B.A., 1997. Cell-cell communication in gram-positive bacteria. Annual review of microbiology, 51(1), pp.527-564.

Fuqua, W.C., Winans, S.C. and Greenberg, E.P., 1994. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. Journal of

bacteriology, 176(2), pp.269-275.

Gündoğdu, S., 2023. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. Multimedia Tools and Applications, pp.1-19.

Horswill, A.R., Stoodley, P., Stewart, P.S. and Parsek, M.R., 2007. The effect of the chemical, biological, and physical environment on quorum sensing in structured microbial communities. Analytical and bioanalytical chemistry, 387, pp.371-380.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P. and Carman, M., 2016. Are word embedding-based features useful for sarcasm detection?. arXiv preprint arXiv:1610.00883.

Le, N.Q.K. and Huynh, T.T., 2019. Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation. Frontiers in Physiology, 10, p.1501.

Lv, H., Dao, F.Y., Zulfiqar, H. and Lin, H., 2021. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. Briefings in Bioinformatics, 22(6), p.bbab244.

Liu, C.M., Ta, V.D., Le, N.Q.K., Tadesse, D.A. and Shi, C., 2022. Deep neural network framework based on word embedding for protein Glutarylation sites prediction. Life, 12(8), p.1213.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X. and Alzheimer's Disease Neuroimaging Initiative, 2018. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Frontiers in neuroscience, 12, p.777.

Mangwani, N., Dash, H.R., Chauhan, A. and Das, S., 2012. Bacterial quorum sensing: functional features and potential applications in biotechnology. Journal of molecular microbiology and biotechnology, 22(4), pp.215-227.

MB, M., 2001. Bassler BL. Quorum sensing in bacteria. Annu Rev Microbiol, 55, pp.165-99.

Mohamed, A.E., 2017. Comparative study of four supervised machine learning techniques for classification. International Journal of Applied, 7(2), pp.1-15.

Mostafa, F.A., Afify, Y.M., Ismail, R.M. and Badr, N.L., 2022. Deep learning model for protein disease classification. Current Bioinformatics, 17(3), pp.245-253.

Naili, M., Chaibi, A.H. and Ghezala, H.H.B., 2017. Comparative study of word embedding methods in topic segmentation. Procedia computer science, 112, pp.340-349.

Online accessed: 4 October 2023 [https://biopython.org/wiki/ProtParam?fbclid=IwAR1JWQK34HyW30afjY2bGLZzkq900sPU019z7KZVQtj1ocfk_v16JPBoKFI]

Patrick, M.T., Raja, K., Miller, K., Sotzen, J., Gudjonsson, J.E., Elder, J.T. and Tsoi, L.C., 2019. Drug repurposing prediction for immune-mediated cutaneous diseases using a word-embedding–based machine learning approach. Journal of Investigative Dermatology, 139(3), pp.683-691.

Pu, Y., Li, J., Tang, J. and Guo, F., 2021. DeepFusionDTA: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(5), pp.2760-2769.

Rajput, A., Gupta, A.K. and Kumar, M., 2015. Prediction and analysis of quorum sensing peptides based on sequence features. PLoS One, 10(3), p.e0120066.

Rutherford, S.T. and Bassler, B.L., 2012. Bacterial quorum sensing: its role in virulence and possibilities for its control. Cold Spring Harbor perspectives in medicine, 2(11), p.a012427.

Sandhiya, S. and Palani, U., 2020. An effective disease prediction system using incremental feature selection and temporal convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 11(11), pp.5547-5560.

Sarkar, D., Poddar, K., Verma, N., Biswas, S. and Sarkar, A., 2020. Bacterial quorum sensing in environmental biotechnology: a new approach for the detection and remediation of emerging pollutants. In Emerging Technologies in Environmental Bioremediation

(pp. 151-164). Elsevier.

Selva Birunda, S. and Kanniga Devi, R., 2021. A review on word embedding techniques for text classification. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, pp.267-281.

Shi, L. and Chen, B., 2021, August. LSHvec: a vector representation of DNA sequences using locality sensitive hashing and FastText word embeddings. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 1-10).

Sifri, C.D., 2008. Quorum sensing: bacteria talk sense. Clinical infectious diseases, 47(8), pp.1070-1076.

Sivaramakrishnan, M., Suresh, R. and Ponraj, K., 2022. Predicting quorum sensing peptides using stacked generalization ensemble with gradient boosting based feature selection. Journal of Microbiology, 60(7), pp.756-765.

Skandamis, P.N. and Nychas, G.J.E., 2012. Quorum sensing in the context of food microbiology. Applied and Environmental Microbiology, 78(16), pp.5473-5482.

Tékouabou, S.C., Gherghina, Ș.C., Toulni, H., Mata, P.N. and Martins, J.M., 2022. Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. Mathematics, 10(14), p.2379.

Wang, C., Wu, J., Xu, L. and Zou, Q., 2020. NonClasGP-Pred: robust and efficient prediction of non-classically secreted proteins by integrating subset-specific optimal models of imbalanced data. Microbial genomics, 6(12).

Waters, C.M. and Bassler, B.L., 2005. Quorum sensing: cell-to-cell communication in bacteria. Annu. Rev. Cell Dev. Biol., 21, pp.319-346.

Wei, L., Hu, J., Li, F., Song, J., Su, R. and Zou, Q., 2020. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Briefings in Bioinformatics, 21(1), pp.106-119.

Wu, S., Liu, J., Liu, C., Yang, A. and Qiao, J., 2020. Quorum sensing for population-level

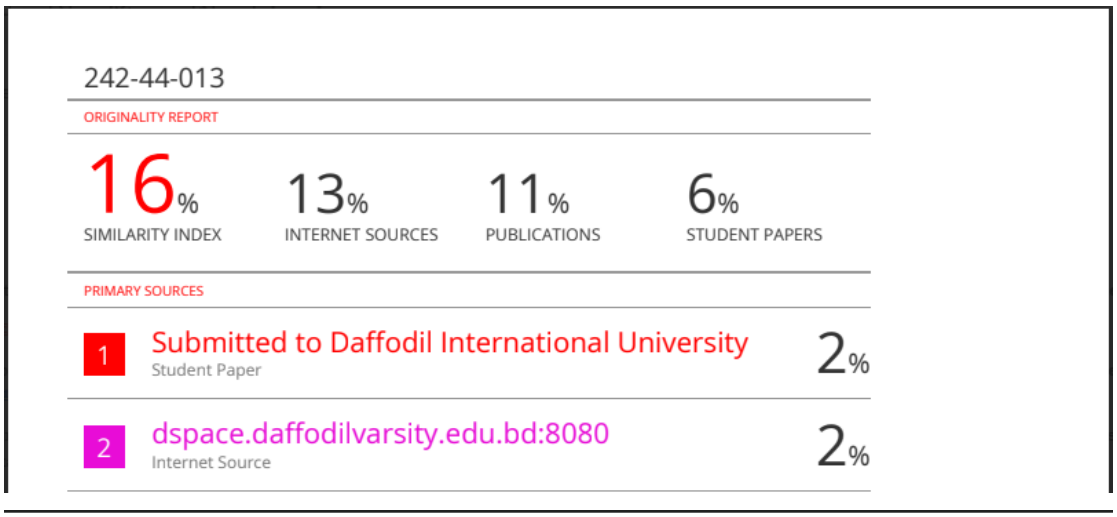
control of bacteria and potential therapeutic applications. Cellular and Molecular Life Sciences, 77, pp.1319-1343.

CHAPTER 8: APPENDIX

Short Form	Abbreviations
QSP	Quorum Sensing Peptides
LSA	Latent Semantic Analysis
PAAC	Pseudo Amino Acid Composition
CNN	Convolutional Neural Network
MCC	Matthews correlation coefficient
QS	Quorum sensing
ML	Machine Learning
AAC	Amino Acid Composition
ACRP	Amino Acid Residue Position
PCP	Physicochemical Properties
GRAVY	Grand average of hydropathy
SVM	Support Vector Machine
RF	Random Forest
DPC	Dipeptide Composition
DEM	Dipeptide Deviation from Expected Mean
TPC	Tripeptide Composition
DL	Deep Learning
CV	Cross-Validation
CTDT	Composition, Transition, Distribution, Transition
GDPC	Grouped Dipeptide Composition

SVD	Singular Value Decomposition
GB	Gradient Boosting
ADB	Adaptive Boosting
XGB	Extreme Gradient Boosting
LGBM	Light Gradient Boosting Machine

PLAGIARISM REPORT



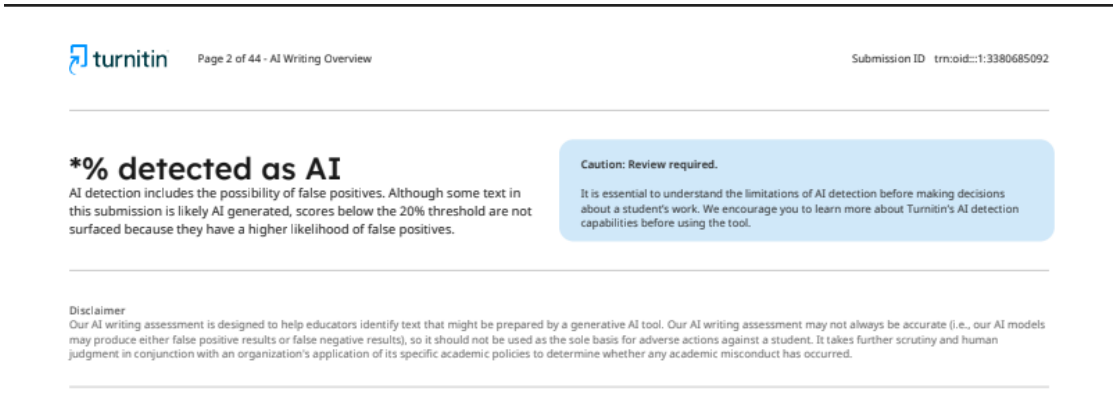
242-44-013

ORIGINALITY REPORT

16%	13%	11%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%



turnitin Page 2 of 44 - AI Writing Overview Submission ID trn:oid:::1:3380685092

***% detected as AI**

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.