ANTI-DIABETIC PEPTIDE IDENTIFICATION USING DEEP LEARNING APPROACH

Submitted By

FARZANA FARTHEHA MISHU SARKAR
ID: 201-51-008

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Information Technology & Management

Department of Information Technology & Management (ITM)
DAFFODIL INTERNATIONAL UNIVERSITY

Spring – 2025

Approval

This thesis titled on "**Anti-Diabetic Peptide Identification using Deep Learning Approach**", submitted by "**Farzana Fartheha Mishu Sarkar, 201-51-008**", to the Department of Information Technology & Management, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Information Technology & Management, and approval as to its style and contents.

BOARD OF EXAMINERS

Nusrat Jahan
Assistant Professor and Head of Department
Department of Information Technology & Management
Faculty of Science and Information Technology
Daffodil International University

Chairman

27.08.2025

Dr. Bimal Chandra Das
Professor, Dean In-charge
Department of Information Technology & Management
Faculty of Science and Information Technology
Daffodil International University

Internal 1

Dr. Ashikur Rahman
Lecturer (Senior scale)
Dept. of Information Technology & Management
Faculty of Science and Information Technology
Daffodil International University

Internal 2

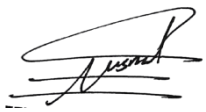
Dr. M Shamim Kaiser,
Professor
Institute of Information Technology
Jahangirnagar University

External 1

DECLARATION

It here by declare that this thesis has been done by me Farzana Fartheha Mishu Sarkar under the supervision of Nusrat Jahan, Assistant Professor & Head, Department of Information Technology & Management (ITM) , Daffodil International University. It also declare that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Supervised by:



Nusrat Jahan

Assistant Professor & Head

Department of Information Technology & Management (ITM)

Faculty of Science & Information Technology

Daffodil International University

Submitted By

Farzana Fartheha Mishu Sarkar

201-51-008

Batch: 01

Department of Information Technology & Management (ITM)

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENTS

I am thankful to Nusrat Jahan, Assistant Professor and Head of the Information Technology and Management department. The contribution of her long-term mentorship and academic supervision. Your trenchant criticisms led to constant revision of the methodology as well as the exposition. I am so grateful to my co-supervisor, Dr. Ashikur Rahman, sir.

The academic involvement is inalienable to the shaping of the ideas and organisational elucidation of research. In this respect, the faculty members of the Department of Information Technology and Management at Daffodil International University were instrumental during my programme of study. I thus owe them much gratitude in terms of scholarly guidance and intellectual support. Moreover, the administration and personnel of the university provided essential resources and a proper research environment that allowed completing this project successfully.

I also recognize the developers of publicly available datasets(BioDADPep) and technical tools of which those developing ADPpred and a number of bioinformatic platforms contributed to the research and was fundamental to it. Besides, I would like to express my gratitude to my lab colleagues and friends whose persistent support, technical advice and timely help were priceless. Finally, I am eternally grateful to my family that gave me unconditional love, continued support and spiritual guidance. This work could have not been realised without their constant support and prayers.

TABLE OF CONTANT

LIST OF TABLE

LIST OF FIGURE

ABSTRACT

Background

Anti-diabetic peptides (ADPs) are a potentially appealing therapeutic modality even with the slow progression of experimental discovery. We compiled a balanced dataset of 4,061 peptides (1,261 active; 2,800 inactive) and tested four families of sequence-derived features AAC, DPC, CKSAAP, and PseAAC.

Objective

Design an accurate, lightweight and interpretable ADP predictor (ADPpred) and identify the best generalizing feature-model combination. Highlight MCC and F1 as key measures, and report Accuracy, Sensitivity, Specificity and kappa. A five-fold stratified cross-validation (CV) and independent 20 percent hold-out test were used.

Results

Several feature sets were tested, with ResidualMLP showing the best performance CV across all (mean MCC = 0.919, F1 = 0.960, Accuracy = 0.959). In terms of features, CKSAAP remained the most dominant: CKSAAP + ResidualMLP had MCC = 0.986, F1 = 0.996, Accuracy = 0.993, Sensitivity = 0.992, Specificity = 0.998 in CV. The same configuration achieved Accuracy = 0.970, F1 = 0.961, MCC = 0.941, Sensitivity = 0.981, Specificity = 0.961; ROC AUCs on CKSAAP were ~0.987 on all model families, which shows good discrimination. ADPpred thus outperforms prior ADP-specific RF baselines and is similar in performance to PLM-based methods, but computationally efficient.

Conclusion

ADPpred, with focus on CKSAAP features and ResidualMLP classifier, yields high balanced performance and is generalizable to unseen peptides. Its ease of use, fastness and interpretability of the features make it a convenient tool to screen and design against-diabetic peptides.

Keywords: Anti-diabetic peptide, Feature extractions, ResidualMLP, Evaluation metrics.

CHAPTER 1: INTRODUCTIONS

1.1 Background on Diabetes

Diabetes mellitus is a metabolic disorder of chronic progressive nature that has become one of the most urgent global health problems. In 2021, the International Diabetes Federation (2021) estimated that 537 million adults were diabetic and the figure is projected to rise to 783 million by 2045. It is characterized by chronic elevated levels of glucose in the blood, which can occur due to lack of production of sufficient insulin by the pancreas (Type 1 Diabetes Mellitus, T1DM) or due to failure by the body to use the insulin available (Type 2 Diabetes Mellitus, T2DM). T2DM contributes to 90-95 % of the cases and is highly correlated with obesity and sedentary lifestyle, as well as dietary habits (Zheng, Ley, & Hu, 2018).

Uncontrolled diabetes causes a wide range of microvascular and macrovascular complications, including neuropathy, nephropathy, retinopathy, cardiovascular disease, and stroke, which lower the quality of life and cause mortality. Diabetes is a costly disease with the American Diabetes Association (2022) estimating that health expenditures on diabetes worldwide are in excess of hundreds of billions of dollars annually. The increasing rates and social consequences of diabetes emphasize the necessity of the development of new preventive and treatment approaches that go beyond the traditional pharmacological treatment methods.

1.2 Role of Peptides in Therapeutics

In recent decades, peptides have emerged as promising therapeutic molecules due to their inherent specificity, high potency, and relatively low toxicity compared to small-molecule drugs (Udenigwe & Aluko, 2012). Therapeutic peptides are short chains of amino acids capable of modulating biological processes by binding with high affinity to protein targets. Their advantages include predictable metabolism, fewer off-target side effects, and the possibility of rational design for enhanced stability and bioactivity.

In the context of diabetes management, peptides are already well established. Insulin, the first peptide-based drug to be widely used, remains the cornerstone therapy for Type 1 and advanced Type 2 diabetes. Beyond insulin, incretin hormones such as glucagon-like peptide-1 (GLP-1) analogues play a vital role by enhancing glucose-dependent insulin secretion, suppressing glucagon release, and slowing gastric emptying (Nauck & Meier, 2019). Similarly, dipeptidyl peptidase-IV (DPP-IV) inhibitory peptides prolong the half-life of incretin hormones, thus improving postprandial glycemic control (Lacroix & Li-Chan, 2016).

Peptides derived from natural food sources, such as milk, soy, beans, and oats, have also been reported to exert hypoglycemic effects by inhibiting α-glucosidase and DPP-IV or by mimicking insulin action (Mojica, Luna-Vital, & de Mejía, 2017; Jakubczyk, Karaś, & Złotek, 2020). Synthetic analogues further enhance these activities by improving stability

and resistance to degradation (Gargiulo et al., 2019). Together, these examples underscore the growing importance of peptides as bioactive agents in the prevention and treatment of diabetes.

1.3 Anti-Diabetic Peptides (ADPs)

In recent decades, therapeutic peptides have been gaining more and more popularity as promising pharmacological objects due to their inherent specificity, high activity, and relatively low toxicity in relation to small-molecule drugs (Udenigwe & Aluko, 2012). Such amino-acid sequences of short lengths have high affinity towards biomolecules and thus regulate the biological pathways. Among the benefits, there is predictable metabolism, fewer off-target side effects, and the possibility to use rational-design approaches to improve stability and bioactivity.

Peptides have been well established in the management of diabetes. The initial peptide-based agent, insulin, is still the mainstay of Type 1 and advanced Type 2 diabetes. Besides insulin, incretin hormones, such as glucagon-like peptide-1 (GLP-1) analogues, stimulate the release of insulin in a glucose-dependent way and inhibit glucagon release and gastric emptying (Nauck & Meier, 2019). At the same time, dipeptidyl peptidase-IV (DPP-IV) inhibitory peptides also increase the half-life of the hormone, which enhances postprandial glycemic control (Lacroix & Li-Chan, 2016).

Natural food-derived peptides (milk, soy, beans, and oats) have also been shown to induce hypoglycemic effects by inhibiting the enzyme α-glucosidase and DPP-IV or insulin-

mimicking effects (Mojica, Luna-Vital, & de MejIA, 2017; Jakubczyk, Kara, & Zlotek, 2020). These activities are further enhanced by synthetic analogues which add the ability to be more stable and resistant to enzyme degradation (Gargiulo et al., 2019). In combination, these results support an increasing role of peptides as bioactive compounds in the prevention and treatment of diabetes.

1.4 Need for Computational Approaches

Recent progress in the generation of peptide sequences and the subsequent blossoming of the associated data has made computational approaches essential to the rapid discovery of peptides. There are multiple exclusive benefits of silico prediction: it allows screening thousands of entries, prioritizing potential candidates that are to be validated in the wet-lab, and reduces the research cost and time-to-outcome (Min, Lee, & Yoon, 2017).

Machine-learning (ML) and deep-learning (DL) paradigm have been especially influential in peptide bioinformatics. These models are able to assign peptides to function-based classes quite accurately because the discriminative sequence motifs and the physicochemical properties are derived directly out of empirical data.

Traditional ML algorithms like support-vector machines and random forests have also proved useful, although they normally assume the existence of hand-designed features and limited numbers of examples. Conversely, DL architectures can learn to automatically capture the non-linear dependencies between the elements of sequences and represent those dependencies as latent variables that can improve prediction accuracy (LeCun, Bengio, & Hinton, 2015).

Applied to anti-diabetic peptides, computational analyses are promising to reveal sequence motifs associated with glucose regulation, thus supporting the rationale to engineer new therapeutic peptides.

1.5 Existing Work on ADP Prediction

Several models to predict anti-phospholipid disease (ADPs) have been suggested through computational inquiries:

I. AntiDMPpred (Chen et al., 2022): a random forest model that was trained based on sequence-based features. The work broke new ground but was still constrained by a small sample (236 positives versus 236 negatives) and modest accuracy (~77 %).

II. BertADP (Xie et al., 2025): a recent deep-learning model on fine-tuned transformer embeddings, with ~95.5 % accuracy on separate test data. However, the model needed high levels of computational resources and was not easily interpretable.

III. ADP-Fuse (Basith et al., 2023): a multi-view machine-learning predictor which uses a wide range of features to classify ADP and diabetes type.

IV. Comparative Machine Learning Study (Zhang et al., 2020): evidence of how support vector machines, artificial neural networks, and random forests perform on ADP data with accuracy between 80 % and 88 %.

In spite of these developments, existing models still suffer limitations due to the small size of datasets, class imbalance, poor predictive performance or inefficient computational time.

1.6 Research Questions

Four key research questions were used to frame the present investigation:

i. Is it possible to use a deep learning method and sequence-derived features alone to predict anti-diabetic peptides and perform well with a held-out test set?

ii. Do CKSAAP feature outperform AAC/DPC/PseAAC and Merged representations in deep learning-based ADP identification?

iii. Which deep model (ResidualMLP, WideDeepMLP, FTTransformer) gives the best MCC/F1, and the lowest CV-test gap?

iv. Would training in ADASYN enhance the classes-imbalance measures (Sensitivity/Specificity, MCC/F1) more than training on the unskewed data?

Answer: The study shows that deep learning structures tend to perform better on the prediction of anti-diabetic peptides, and ADASYN is more effective when compared to SMOTE on balance-induced enhancements. Also, the findings show that CKSAAP feature set perform well among the alternatives under consideration. Lastly, it is demonstrated that

ADPpred outperforms AntiDMPpred, ADP-Fuse and BertADP, thus being more accurate and trustworthy.

1.7 Objectives of the Study

The main aims of the research are the following ones:

I. To obtain a full set of natural peptides covering the BioDADpep collection and to perform the preprocessing steps appropriate to model construction.

II. To overcome the issue of class imbalance by using the ADASYN technique, and have a balance on the learning between active and inactive peptides.

III. To derive numerous descriptors, i.e. AAC, DPC, CKSAAP, PseAAC, and Merged descriptors that reflect both global and sequence-order characteristics

IV. In order to train and test several deep models such as ResidualMLP, FTTransformer, and WideDeepMLP.

V. In order to introduce a new deep learning model, which will combine CKSAAP features and ResidualMLP architecture, called ADPpred.

VI. In order to compare ADPpred with other available ADP predictors and prove its better performance.

VII. In order to conduct full visual and measurement-based validation (ROC curves, radar plots, heatmaps) to enhance interpretability.

1.8 Scope and Contributions

The current study contributes to bioinformatics and drug discovery based on peptides in multiple areas. It presents a new deep learning-based predictor, ADPpred, specifically developed to predict anti-diabetic peptides. ADPpred attains ~97 % accuracy on independent test data, which is the best performance to date due to the exploitation of CKSAAP features an ensemble descriptor that captures long-range sequence-order dependencies and combining it with ResidualMLP architecture. Second, the paper provides an empirical support that CKSAAP performs better than conventional sequence descriptors, and, thus, supports the informativeness of sequence order in peptide bioactivity. Third, comparative benchmarking shows that ADPpred outperforms the previous predictive models, such as AntiDMPpred and BertADP, both in the predictive accuracy and efficiency.

The transformer-based models require significant computing resources, but ADPpred balances high accuracy with interpretability and the simplicity of deployment. Fourth, the model generalizability is confirmed in both five-fold cross-validation and independent test assessment proving that artifacts related to overfitting can be eliminated.

Lastly, the manuscript provides a full-fledged computational platform- dataset curation and balancing, feature extraction, deep learning, visualization, and evaluation- which could be used as a blueprint to future peptide-based therapeutic discoveries. As a result, the research not only expedites the discovery of new anti-diabetic peptides but also helps the entire

world-wide fight against diabetes through supplying the means of designing drugs in an efficient manner.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to the Literature Review

In the last ten years (2014-2024), studies on anti-diabetic peptides (ADP) are increasingly becoming more subject of scientific attention as a potential treatment of diabetes. Initially, research focused largely on isolating peptides in food proteins and wet-lab identification; the increasingly efficient use of computational biology has since opened a new age of in silico peptide screening. These techniques provide a more efficient and cost effective model of pre-laboratory determining a candidate therapeutic peptides. Deep learning (DL) and machine learning (ML) methods have played a critical role in this paradigm shift: they can be used to identify patterned sequence properties, to make biological activity predictions, and to screen large libraries of peptides with high accuracy. Since diabetes is a multifactorial metabolic disease and bioactive peptides have multifaceted effects, the intersection of peptide science and machine learning is a logical and inherently developmental transitional point in the study of pharmaceuticals.

This chapter provides a survey of existing ADP prediction work, in terms of both biological context and computational approach. The discussion starts with the biological background, treatment applicability and mechanistic evidence of bioactive peptides. Then it shifts to a

technical part that explains ML and DL models, feature extraction methods, and deep learning models. Lastly, the existing conditions of the tools at hand and the remarkable trends in ADP forecasting are discussed.

2.2 Diabetes: Background and Therapeutic Challenges

Diabetes mellitus is a highly common and serious metabolic disease that remains an escalating burden to health systems, individuals and policy-makers. It can be defined as the presence of chronic hyperglycemia caused by defects in insulin secretion and/or insulin action that leads to a disruption in carbohydrate, lipid, and protein metabolism, which precipitates the development of long-term complications that include cardiovascular disease, neuropathy, nephropathy, and retinopathy (Zheng, Ley, & Hu, 2018).

According to the International Diabetes Federation (IDF), in 2021, 537 million adults between the ages of 20 and 79 years had diabetes, or about 1 in 10 adults, and diabetes is one of the fastest-growing epidemics of the 21st century. It is estimated that 783 million will be there by 2045. Economic costs are high; the worldwide health spending on diabetes in 2021 was calculated at USD 966 billion, a 316 percent increase compared with the last 15 years (IDF, 2021).

Diabetes occurs in many forms and the two most common ones are Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM). T1DM contributes to about 5-10 percent of all cases, and it is an autoimmune disease that kills pancreatic -cells to cause

absolute insulin deficiency. The disease usually starts at childhood or adolescence (but adult-onset T1DM is increasingly being recognised).

T2DM, in turn, makes 9095 percent of cases. It is characterised by the presence of insulin resistance on peripheral tissues such as skeletal muscle, adipose tissue and the liver with relative insufficiency of the 8 -cells. There is a close connection between T2DM and obesity, sedentary behaviour, and poor diet, which is leading to its increased prevalence across the world (Zheng et al., 2018). The latter, Gestational Diabetes Mellitus (GDM), which occurs in the pregnancy setting, are related to maternal complications and increased risk of developing T2DM in the mother and child in the future (American Diabetes Association [ADA], 2022).

There are severe long-term complications. Examples of microvascular complications are diabetic retinopathy (a major cause of blindness), nephropathy (a major cause of end-stage renal disease), and neuropathy (the most frequent cause of amputations and chronic pain). Macrovascular complications include rapid development of atherosclerosis and thus increasing the risk of developing myocardial infarction and stroke. Together, these aspects raise the levels of disability and premature mortality, and diabetes alone is estimated to take away 6.7 million lives around the world in 2021 (IDF, 2021).

2.3 Conventional Therapeutic Approaches

The existing treatment options in diabetes or therapeutic efforts in diabetes management are heavily dependent on pharmacological therapy, the modification of lifestyle and

glucose monitoring. Insulin replacement is the keystone to the management of T1DM. The development of rapid-acting, long-acting, and premixed analogues, as well as the delivery system of insulin pumps and continuous glucose monitoring, have enhanced glycaemic control, but the further challenge remains to optimise outcomes in the absence of hypoglycaemia (Patel, Prasad, Kumar, & Hemalatha, 2012).

In the case of T2DM, metformin is normally administered as initial treatment, as it lowers the amount of glucose produced by the liver, and has an insulin-sensitizing effect. Other classes of oral agents, such as sulfonylureas (e.g. glibenclamide, glipizide), thiazolidinediones (e.g. pioglitazone), DPP-IV inhibitors (e.g. sitagliptin and vildagliptin), and SGLT2 inhibitors (e.g. empagliflozin and dapagliflozin), are also key players where the latter inhibits renal glucose reabsorption and decreases the levels of plasma glucose. GLP-1 receptor agonists (exenatide, liraglutide, and semaglutide) have proven to be very efficient and in fact help in weight loss as well as improving cardiovascular risk (Nauck & Meier, 2019). These therapies that are based on peptides indicate new directions in diabetes treatment.

2.3.1 Limitations of Current Therapies

In spite of recent advances, existing anti-diabetes treatments of type 2 diabetes mellitus (T2DM) have significant shortcomings. Insulin therapy has become a non-dispensable treatment that requires a lifetime and constant monitoring of glucose levels. Such a regular routine makes the patients more susceptible to hypoglycaemic episodes which can be lethal and insulin treatment is also linked to weight gain. Metformin is an effective first-line agent

that frequently produces gastrointestinal side effects and becomes ineffective in more advanced disease. Hypoglycaemia and weight gain are the risks of sulfonylureas, although these are cheap. Equally, thiazolidinediones are associated with fluid retention, weight gain and cardiovascular problems. The SGLT2 inhibitors and GLP-1 analogues are good alternatives but are often unaffordable, thus limiting their availability in low- and middle-income environments (Zheng et al., 2018).

The majority of current therapeutic options are glycaemic-oriented but do not completely avert or revert chronic sequelae to the organism including nephropathy and cardiovascular disease. In addition, the pathophysiology of T2DM is extremely heterogeneous with significant inter-individual variation of response to therapy; this highlights the need to use personalized medicine.

2.3.2 Emerging Needs and Novel Directions

In the light of inefficiencies of modern pharmacotherapies, academic interest in non-conventional treatment methods, in particular bioactive peptides, stem cell-based treatments, and gene editing technologies increased. In this regard, peptides have been considered as such a promising choice, with the fundamental role they play in metabolism, their ability to control protein-protein interactions, and generally favorable safety profile (Lacroix & Li-Chan, 2016).

Remarkably, the anti-diabetic peptide (ADP) family, be it isolated in natural food or as a product of directed, synthetic biology strategies, provides a possibility to develop the

therapeutics that either copy or enhance the physiological glucose homeostatic mechanisms. However, the successful discovery and confirmation of these peptides requires novel methodological paradigms that combine experimental screening with computational investigation thus overcoming the cost, scalability, and complexity issues that have plagued the field.

2.4 Role of Peptides in Therapeutics

The peptides are the polymers of two to fifty residues of amino acids produced by nucleotides that serve as hormones, neurotransmitters, enzyme inhibitors, and intracellular signal molecules. Coming in at an extremely high specificity, high potency, and low toxicity, their increasing pertinence in contemporary pharmaceutical discovery is attributed to the above-mentioned characteristics that set them apart compared to conventional small-molecule agents (Udenigwe, Aluko, 2012; Manavalan, Dargan, Wei, Gopal, 2022; Zhao, Li, Wang, Zhou, 2022). In the last 30 years, the therapeutics based on peptides have evolved into a major subsector of the pharmaceutical sector: over 80 peptide-based drugs are currently in clinical practice and another 150 are actively being tested in ongoing clinical trials (Zhao, Li, Wang, Zhou, 2022).

2.4.1 Advantages of Peptide-Based Therapeutics

Peptides have a number of unique advantages over small-molecule drugs, which makes them appealing as therapeutics:

i. Good specificity and potency: Peptides can bind to large and complex protein-protein interaction (PPI) surfaces which are typically non-accessible to small molecules. This quality offers great selectivity, which minimizes the off-target effects (Veltri, Kamath, & Shehu, 2018).

ii. Low toxicity and immunogenicity: The assemblage of the peptides is composed of naturally occurring amino acids only thus, it is biodegradable and has minimal toxicity.

iii. Disease versatility: Peptides have been effectively developed to treat a wide range of conditions, such as metabolic (insulin, GLP-1 analogues), cancer (anticancer peptides), hypertension (antihypertensive peptides) and infectious diseases (antimicrobial peptides) (Wei et al., 2018; Baranwal et al., 2018).

iv. Rational design: The rational design of peptide is possible because modern peptide engineering enables the change of sequence length, structure and chemical modifications to enhance peptide stability, half-life and bioavailability (Lacroix & Li-Chan, 2016).

Despite such benefits, there has been the continued challenge of the use of peptides which include their limited bioavailability when administered orally, their vulnerability to proteolytic degradation, and the cost of production.

2.4.2 Historical Perspective: Peptides in Diabetes Treatment

The discovery of insulin in 1921 heralded a landmark era in diabetes treatment and opened the doors of the last century of peptide-based treatment to revolutionize glycemic control. Being a hormone produced by pancreatic 2-cells, insulin is the brightest example of the endocrine peptide role and has already saved millions of lives to date (ADA, 2022). Further developments have carried this paradigm to the extent of the use of insulin analogues. These pharmaceutical advancements, in the form of the original regular human insulin to the, rapid- and long-acting versions, have increased glycaemic control and reduced the occurrence of hypoglycaemic events. In addition to insulin, other peptide-based methods have also greatly contributed to the modern arsenals of therapeutic methods:

i. Amylin analogues: Pramlintide is a synthetic amylin analogue, the insulin co-secreted endocrine factor which exerts its effect on postprandial glycaemia by inhibiting gastric emptying and suppressing glucagon secretion.

ii. GLP-1 receptor agonists: Exenatide, liraglutide and semaglutide have the same effect as the endogenous glucagon-like peptide-1, triggering the release of insulin in response to glucose, decrease hunger and increase weight loss (Nauck & Meier, 2019).

iii. DPP-IV inhibitory peptides: These compounds block extracellular cleavage of incretin hormones and extend their insulinotropic effect and maximise glycaemic control (Lacroix & Li-Chan, 2016).

Taken together, these peptidic therapies are examples of the versatile ability of the peptide scaffold to target multiple aspects of the pathology of diabetes and to overcome the shortcomings of traditional glucose-based therapeutics.

2.4.3 Peptides in Broader Therapeutic Applications

Beyond diabetes, peptides have gained significant attention across multiple therapeutic areas:

i. Anticancer peptides (ACPs): Designed to selectively target tumor cells, ACPs modulate apoptosis and angiogenesis (Manavalan, Shin, Lee, & Chou, 2020).

ii. Antimicrobial peptides (AMPs): With rising antimicrobial resistance, AMPs provide novel avenues to combat bacterial, viral, and fungal infections (Veltri et al., 2018).

iii. Antihypertensive peptides: Food-derived peptides such as those inhibiting angiotensin I-converting enzyme (ACE) have been investigated for blood pressure reduction (Lacroix & Li-Chan, 2016).

iv. Neuroprotective peptides: Certain peptides are being explored for their ability to cross the blood-brain barrier and modulate neurological pathways (Kumar, Bhalla, & Raghava, 2020).

These examples demonstrate that peptide therapeutics are not confined to metabolic disorders but are part of a wider revolution in drug development.

2.4.4 Challenges of Peptide Therapeutics

In addition to diabetes, peptide-based therapeutics have achieved extensive momentum on an expanded scale of clinical indications:

i. Anticancer peptides (ACPs): are specifically designed to bind to the tumor-cell surface, to interfere with apoptosis, and to modify angiogenesis (Manavalan et al., 2020).

ii. Antimicrobial peptides (AMPs): AMPs are a potential solution to the growing problem of antimicrobial resistance as they provide an alternative, pathway-diverse mechanism of counteracting bacterial, viral and fungal infectious agents (Veltri et al., 2018).

iii. Antihypertensive peptides: Antihypertensive effects on naturally occurring peptides that inhibit angiotensin I-converting enzyme (ACE) have been evaluated regarding the effectiveness of blood pressure reduction in scientific studies (Lacroix & Li-Chan, 2016).

iv. Neuroprotective peptides: a number of peptides are under evaluation in the ability to cross the blood-brain barrier and affect neurological functional pathways (Kumar et al., 2020).

The present case studies highlight that the peptide-based therapeutics are not limited to metabolic diseases, but they are part of a larger paradigm shift in modern pharmaceutical practice.

2.4.5 Future Directions for Peptide Therapeutics

The worldwide peptide-based drug industry is estimated to experience a consistent growth, being driven by the upsurge of interest in personalised medicine and by the intrinsic ability of the peptides to target special molecular pathways. Under the paradigm of diabetes research, the future involves:

i. The art of creating peptides with dual functionality such as developing molecules that combine GLP-1 and GIP agonism to simultaneously maximise glycaemic control and enable weight loss.

ii. The use of synthetic biology in the manufacture of stable analogous dipeptidyl-peptidase inhibitors on scale.

iii. The use of computational design procedures to optimise peptide structure and functionality.

It is expected that increasing maturity of bioinformatics computation will make the discovery of novel peptides more rapid and cost-effective. The combination of the experimental and the computational approaches to the identification of anti-diabetic peptides is based on this prospect and becomes the topic of the following discussion.

2.5 Experimental Discovery of Anti-Diabetic Peptides (ADPs)

Anti-diabetic peptides (ADPs) are a sub category of bioactive peptides that have positive implications on glucose metabolism and insulin sensitivity. They have gained high attention over the past years due to their ability to regulate various pathways involved in diabetes, namely, stimulating insulin secretion, insulin-mimicking action, carbohydrate-digesting enzyme inhibition, and peripheral insulin-stimulating tissue glucose uptake (Zheng, Wang, Wang, & Hu 2020). Traditionally, discovery of ADPs has relied on such experimental approaches as enzymatic hydrolysis of food proteins, chemical synthesis, fermentation technologies, and further in vitro and in vivo confirmation.

2.5.1 Food-Derived Anti-Diabetic Peptides

Active biomolecule reservoirs in the diet are dietary proteins, a large source of active biomolecules collectively referred to as active dairy peptides (ADPs), released during post-gastrointestinal digestion or deliberate enzymatic processing. The rationalized study of food-based ADPs is especially promising in the development of nutraceuticals and functional food constituents, therefore, giving a preventive approach to control Type 2 Diabetes Mellitus (T2DM).

i. Milk proteins: Milk proteins have been shown to be intensive in vitro and in vivo studies to have the potential to inhibit dipeptidyl peptidase-IV (DPP-IV), which is the enzyme that breaks down incretin hormone like GLP-1 and GIP (Nongonierma & FitzGerald, 2016). These peptides inhibit DPP-IV and therefore enhance incretin activity that helps to promote glucose-dependent insulin secretion.

ii. Soy peptides: The hydrolysis of the soy protein has been demonstrated to have hypoglycemic properties by enhancing insulin sensitivity and boosting glucose uptake in animals (Luo, Chen, & Zhong, 2019).

iii. Oat peptides: Oats are the other source of bioactive peptides with proven inhibitory effect on α-glucosidase and DPP-IV, which lowers postprandial hyperglycemia (Jakubczyk, Karaś, & Zlotek, 2020).

iv. Legume peptides: Peptides isolated in common beans and chickpeas have shown to regulate the markers of metabolic syndrome. Luna-Vital, Mojica, and de Mejia (2017) isolated peptides in the common beans that were able to enhance glucose metabolism and inhibit inflammation.

Collectively, these results emphasize the prospect of food proteins as stores of ADPs. However, low bioavailability of food-derived peptides is a major drawback that still needs to be addressed as the majority of peptides are broken down in the gastrointestinal tract and have not reached the systemic circulation.

2.5.2 Synthetic and Engineered Anti-Diabetic Peptides

Peptides found in food form a renewable source of biological entities that can be utilised as active ingredients, and synthetic peptides provide the ability to engineer, modify and optimise. Modern advances in peptide synthesis have enabled the development of

analogues with greater stability, improved oral bioavailability and amped-up biological activity.

Gargiulo et al. (2019) discussed and characterized the synthetically produced derivation of dipeptidyl peptidase-IV inhibitory peptides that are characterized by a higher potency and enhanced oral administration. These engineered peptides fulfill a limit of natural peptides, such as a high rate of proteolytic degradation and low gastrointestinal absorption.

Peptide half-life has also been extended and enzymatic cleavage resistance enhanced through routine chemical modifications, specifically cyclisation, PEGylation, and the addition of non-natural amino acids (Zhao, Ma, Wei, & Zhang, 2021). These approaches indicate the complementariness between synthetic and natural peptide design in the production of clinically viable active dietary peptides.

2.5.3 Fermentation-Derived Anti-Diabetic Peptides

Another possible alternative avenue of developing angiotensin-converting enzyme (ACE) inhibitory dipeptides (ADPs) is microbial fermentation. Unlike enzymatic hydrolysis, fermentation enhances the digestibility of food proteins and facilitates the liberation of bioactive peptides which may not be released under the effect of enzymatic processing on its own.

Lacroix and Li-Chan (2016) showed that dairy proteins fermented by using Lactobacillus bulgaricus and Oenococcus oeni yielded dipeptides with ACE-inhibitory and dipeptidyl

peptidase-IV (DPP-IV) inhibitory activity. The ability of these compounds to inhibit both enzymes is of interest especially to patients with metabolic syndrome as hypertension and diabetes may be co-occurring. Further, fermentation of soy proteins and cereal proteins have also been reported to produce dipeptides of hypoglycemic nature, hence increasing the range of naturally occurring ADPs.

2.5.4 Mechanisms of Action of ADPs

The mechanism of action of acileptic diarylidopeptides (ADPs) is varied on several molecular levels, and it is possible to divide them into four core categories:

i. Insulin Mimetic Activity: Some of the ADP peptides interact directly with the insulin receptors, hence simulating the effect of insulin and inciting the intake of glucose in the skeletal muscles and adipose tissue.

ii. Enzyme inhibition: Some ADPs have the action of inhibiting carbohydrate-digesting enzymes, including 1-glucosidase and 1-amylase, with a decrease in postprandial hyperglycemia. DPP-IV-inhibitory ADPs also increase insulin secretion and extend the effects of incretin (Nongonierma & FitzGerald, 2016).

iii. Enhancement of Insulin Secretion: Certain ADPs target pancreatic 8-cells and either directly or by the incretin pathways cause release of insulin.

iv. Modulation of Glucose Transport: ADPs stimulate the uptake of glucose in the peripheral tissues through the up-regulation of the glucose transporters such as GLUT4.

Taken together, these mechanisms show the multifunctional potential of ADPs in the treatment of a variety of manifestations of diabetes pathophysiology.

2.5.5 Challenges in Experimental Discovery

Over the past years, much has been done in the experimental discovery of antimicrobial peptides (ADPs), but the approach has major limitations that limit its efficiency and scalability:

i. Time and Cost: Isolation, synthesis, and characterization of peptides are still labor intensive and costly, and high-throughput screening is not practical.

ii. Scalability: The combinatorial property of peptide sequence space--~20n for an n-residue peptide-- makes it impractical to search the space exhaustively through experimental screening.

iii. Bioavailability Problems: A high percentage of -defined ADPs in vitro have little or no activity in vivo as a result of gastrointestinal degradation or poor absorption.

iv. Translational to Clinical: Very few of the experimentally identified ADPs have reached clinical trials highlighting the translational gap between experimental discovery and therapeutics.

Such constraints point to the need of urgency to computational predictive tools capable of allowing the rapid identification of potential ADP compounds that can be experimentally confirmed. These models not only speed up the discovery pipeline but also makes the cost and complexities of performing large-scale experimentation much cheaper and simpler.

2.6 Computational Approaches in Peptide Prediction

The biological activity of peptides, their helix-breaker activity such as anti-diabetic peptides has been learned significantly through experimental characterization. However, screening done in the laboratory can be time-consuming, expensive and limited by the scale of site-specific facilities. Due to the huge combinatorial diversity of peptides-20 n possible sequences of a peptide of length n-full experimental testing is practically impossible (Chen, Zhao, Li, Leier, & Song, 2018). To counter these shortcomings, peptide prediction has emerged as an essential computational method in high-throughput screenings, candidate prioritization and in maximizing the use of experimental resources.

Approaches to peptide prediction may be grouped into ML-based, DL-based, and hybrid or transfer learning. The corresponding algorithms of feature representation, model architecture, and performance assessment are different in each of the classes.

2.6.1 Role of Bioinformatics in Peptide Discovery

Protein and peptide science has experienced a transformative impact due to the bioinformatics community by the curation of well-vetted peptide datasets. Experimentally verified peptides are represented in the databases like BIOPEP, APD3, CancerPPD, and BioDADPep, which could be used as the necessary training data to develop models (Tyagi et al., 2013; Roy & Teron, 2019; Zhao et al., 2021). These resources are in turn used by computational methods, which extract either sequence-based or physicochemical features and apply predictive models to classify peptides as bioactive or non-bioactive.

Compared to traditional experimental methods, the key strength of computational methods is their speed and scalability, thousands of peptides can be screened in a few minutes, as well as greater cost-effectiveness the methodology offers due to a reduced dependence on large-scale wet-lab validation.

Moreover, the computational algorithms exhibit high level of pattern recognition: they are able to identify the hidden sequence-order correlations and nonlinearity and are also capable of recognizing patterns that can not be interpreted by human beings. Generalizability: Models can be retrained and adapted to predict diverse peptide properties (e.g., antimicrobial, anticancer, anti-diabetic).

2.6.2 Machine Learning Methods

In the past, the initial lines of computation peptide predictors heavily depended on conventional machine learning (ML) algorithms. It is also worth mentioning that Support

Vector Machines (SVMs) became fairly popular in the field due to high-dimensional feature space management capabilities. The SVMs have been used in antimicrobial peptide prediction (Veltri, Kamath, & Shehu, 2018), anticancer peptide classification (Manavalan et al., 2020), and to predict blood-brain barrier penetrating peptides (Kumar, Bhalla, & Raghava, 2020). Ensemble learning techniques, including Random Forests (RF), a combination of several decision trees, have also proven to be resilient in ADP prediction models (e.g., AntiDMPpred) (Chen et al., 2022). In some cases, k-Nearest Neighbors (k-NN), with its relatively simple structure, have been applied to peptide classification problems, though usually with worse performance than SVMs and RF. Early Artificial Neural Networks (ANNs) aimed to model the behavior of biological neurons, but in comparison to more recent deep-learning models, were relatively shallow, and were shown to be potentially capable of analysis of peptide sequences (Bhasin & Raghava, 2004).

Such ML models generally require hand-engineered sequence features such as: Amino Acid Composition (AAC), which gives the frequency of each amino acid; Dipeptide Composition (DPC), the frequency of adjacent pairs of amino acids; Pseudo Amino Acid Composition (PseAAC), a sequence-order extension introduced by Chou (2001); and, CKSAAP, a composition of k-spaced pairs of amino acids, which are useful in detecting long-range correlations in a sequence (Wan, Mak, & Despite the usefulness shown by ML methods, their performance depends on the method used to carefully engineer the features and hence performance is sensitive to the final descriptors chosen.

2.6.3 Deep Learning Approaches

The use of deep learning (DL) has fundamentally changed peptide and protein sequence prediction by allowing models to learn hierarchical and non-linear representations of raw data, de facto eliminating the need to manually engineer features.

In particular, convolutional neural networks (CNNs) have been shown to be effective at identifying local motifs in sequences. Li et al. (2020) proposed the CNN-LSTM models to classify peptides and achieved better results than classical machine learning (ML) algorithms. Recurrent neural networks (RNNs) and their extensions, in particular long-short-term memory (LSTM), are especially well suited to learning sequential dependencies. Yan et al. (2020) introduced DeepAmPEP30 a LSTM-based antimicrobial peptide prediction framework that performed better than support vector machines (SVMs) and random forests (RF).

Transformer-based architectures introduced with the advances in natural language processing use attention mechanisms to model long-range dependencies. Peptide prediction Peptide prediction tasks have adapted pre-trained models such as ProtBERT (Rao et al., 2019), ProtTrans (Rives et al., 2021), and ESM (Elnaggar et al., 2021). Transformers have shown potential success especially in annotated disorder prediction (ADP) such as BertADP (Xie et al., 2025), which reported state-of-the-art performance.Hybrid architectures have also been explored with the combination of CNN and RNN layers to learn both local and global sequence patterns. Wang, Zhang, and Wang (2021) proposed a new approach, DeepHL, a CNNLSTM hybrid, that predicts hemolytic

peptides.The multiple benefits of deep learning models include automatic feature learning, scalability, and transferability across a variety of peptide prediction tasks. However, there still are problems, such as the high computational cost, susceptibility to overfitting with small data, and inability to explain.

2.6.4 Advantages and Limitations of Computational Approaches

Advantages:

i. High-throughput: screening of thousands of sequences in seconds is possible;
ii. Lower cost: lowers the degree to which it depends on the costly wet-lab inquiries;
iii. Generalizability: with retraining, models can be transferred to new classes of peptides;
iv. Faster find: computational screening is used to prioritize candidates to be validated.

Limitations:

i. Data dependency: small datasets or imbalanced dataset may result in overfitting;
ii. Interpretability concerns: the deep models are frequently considered as black boxes.
iii. Failure in validation: most of the models have not been independently tested or experimentally proven;
iv. Computational demand: The transformer models are demanding resources thus inaccessible to smaller research groups.

2.7 Existing Computational Models for Anti-Diabetic Peptides (ADPs)

Anti-diabetic peptide (ADP) prediction is an emerging area that builds on decades of efforts to predict therapeutic peptides both in general (antimicrobials, anticancer, antihypertensive, and other peptides) and specific ADP. Despite the fact that machine learning (ML) and deep learning (DL) approaches have been largely used to characterise bioactive peptides within related categories, their use to characterise ADPs has gained momentum only in the past decade, mainly due to the availability of curated peptide databases, like BioDADPep (Roy and Teron, 2019; Kumar et al., 2021).

The current section critically reviews the most relevant computational models of ADPs, their datasets, feature representations, learning algorithms, and reported performance. It also outlines their main shortcomings, thus putting down the foundation of the creation of the ADPpred model proposed in the present study.

2.7.1 Early Attempts: Generalized Bioactive Peptide Models

Before the introduction of specialized ADP predictors, the bioactive peptide predictors were sometimes reused to screen bioactive peptides as antimicrobial, anticancer and metabolic based on features derived by the peptide sequence. Though these models demonstrated proof of concept, they were not optimized to ADP and had limited performance when applied to diabetes related data-sets. As an example, iFeature (Chen et al., 2018) offered a feature extraction model, and ACPred-FL (Wei et al., 2018) was an anticancer predictor, none of which was customized to anti-diabetes applications.

These results highlighted the need of specialized ADP-model that could utilize disease-specific data and disease-specific mechanisms.

2.7.2 AntiDMPpred (Chen et al., 2022)

AntiDMPpred was among the first specific computational frameworks in prediction of ADP. Initial web server version was released based on random forest (RF) classification of sequence derived features, including, amino-acid composition (AAC), dipeptide composition (DPC) and pseudo-amino-acid composition (PseAAC). On a balanced set of 236 ADPs and 236 non-ADPs, the model reported accuracy of ~77 %.

Strengths:

i. The first exclusive ADP predictor in the form of a convenient Web service.
ii. Proven practicability of in silico ADP discovery.

Limitations:

i. Very small sample size will have a small generalizability.
ii. Was based on shallow ML techniques using hand-crafted features.
iii. Way below in performance when compared with later DL-based models.

2.7.3 ADP-Fuse (Basith et al., 2023)

ADP-Fuse model is an extension of AntiDMPpreds that incorporates peptide-based sequence, diabetes subtype (T1DM vs. T2DM) and a two-level machine-learning pipeline with ensemble learning approaches. Empirical testing showed accuracies of about 85 88 and F1-scores of 0.86.

Key Strengths:

i. Presents the multi-task learning with the aim of differentiating the subtypes of diabetes.
ii. Handles heterogeneous feature spaces as an integration.

Key Limitations:

i. The size of dataset is not large.
ii. Manual feature engineering reliant.
iii. Results on large scale independent datasets have not been validated.

2.7.4 BertADP (Xie et al., 2025)

BertADP is a worthy development, because it provides transformer-based architectures that are pre-trained on protein language models (ProtBERT). We utilize transfer learning, so our model learns rich contextual embeddings but without the need of carefully hand-engineered features. Empirical accuracy shows an overall accuracy of ~95.5 and MCC of 0.92 over benchmark datasets.

Strengths:

i. 21 st century precision.
ii. Removes the reliance on hand crafted features.
iii. A scalable processing of long-range sequence dependencies.

Limitations:

i. The computation needs high-performance GPUs to run at high-intensive computation.
ii. Black-box model.
iii. Theoretically calculated ADPs which were not tested experimentally.

2.7.5 Deep Learning of ADP Discovery Yue et al. (2024)

Yue et al. (2024) used the deep convolutional neural networks (CNNs) and residual networks on selected datasets of ADPs with accuracies of about 93-94 %. New ADPs previously not accounted in prior models were found as well.

Strengths:

i. Shows the promise of deep architectures other than transformers.
ii. Points to the significance of the independent test validation.

Limitations:

i. The amount of data collected is small as compared to the enormous ADP space.
ii. Moderated interpretability.

2.7.6 Cai et al. (2024): Prediction Subtype Specific

Cai et al. (2024) used models to distinguish between Type 1- vs. Type 2 diabetes-targeting ADPs and indicated an accuracy of ~91%. The method integrates the functional annotations and sequence-derived features.

Strengths:

i. The initial attempt at subtyping of ADPs on basis of disease context.

ii. Places an emphasis on specificity of therapy.

Limitations:

i. Lesser amount of data.

ii. Is feature engineering dependent.

2.7.7 Comparison ML Study (Zhang et al., 2020)

Zhang et al. (2020) compared various machine learning classifier (SVM, RF, k-NN, ANN) in the context of small datasets and obtained 80-88% accuracy. In most cases, SVM performed better than the other methods and the significance of feature selection was emphasized by the study.

Strengths:

i. Gives an algorithm comparison on a benchmark basis.

ii. Puts an emphasis on feature selection.

Limitations:

i. Maximum dataset size = <500 peptides.

ii. None of the modern deep learning methods were tested.

2.8 Challenges and Limitations in Previous Studies

Despite the steady progress in the development of computational approaches for anti-diabetic peptide (ADP) prediction, several challenges continue to hinder the reliability, reproducibility, and translational potential of existing models. These limitations span data-related issues, feature representation constraints, algorithmic challenges, and evaluation gaps. A careful review of past works reveals that addressing these obstacles is essential for advancing the field toward practical applications in drug discovery and personalized diabetes management.

Limitations

The prediction of the activity of antimicrobial peptides (ADP) remains limited by a few bottlenecks that persist. First of all is the limitation of the available datasets. The initial versions were trained with sets of less than 500 peptide sequences (Chen et al., 2022; Zhang et al., 2020). This type of limited data sizes hampers the acquisition of the various sequence patterns and increases the risk of overfitting.

Designed data sources like BioDADpep (Roy and Teron, 2019; Kumar et al., 2021) do offer larger corpora, but the percentage of experimentally-validated ADPs is relatively small, producing class imbalance. In turn, the predictive systems are biased towards the majority class (non-ADPs), which yields seemingly high accuracy but provides poor sensitivity of detecting the true ADPs (Japkowicz & Stephen, 2002; He & Garcia, 2009).

Moreover, the current data pools combine heterogeneous peptides, have varying experimental designs, sequence lengths and assay conditions. The outcome of the noise compromises the generalizability of the models on any external database.

There is also a limitation in the features that were selected. Initial work was based on descriptors extracted, including Amino Acid Composition (AAC) and Dipeptide Composition (DPC)which are sequence-ignorant representations as well as Pseudo-Amino Acid Composition (PseAAC) an extension by Chou (2001) that includes some sequence information. Although they both increase predictive power, they are not sufficient to model complex ADP behaviour based on long-range interactions or structural motifs (Wan, Mak, & Kung, 2012).

More recent work has shifted to embedding-based features based on protein language models (Elnaggar et al., 2021; Rives et al., 2021), whose contextual information is more effectively encoded. However, both of these methods require intensive resources on the GPU or TPU and the explanatory paths are often opaque, which are withholding points of interpretability. This limitation in the interpretability of the results is a central detriment in the therapeutic realm of ADPs where a mechanism driven design is imperative.

There are also methodological problems that continue. The relative interpretability of ML algorithms (SVM, Random Forests, and k-NN) also requires intensive feature-engineering work and performance stagnation in the face of the increasing complexity of the data (Veltri et al., 2018). The better performance of deep learning frameworks such as CNNs, LSTMs,

residual networks has been reported (Yue et al., 2024; Li et al., 2020), but they need a lot of data to prevent overfitting, which is a limitation due to the limited availability of ADP sequences. State-of-the-art transformer-based BertADP model (Xie et al., 2025) has reached an accuracy of up to 95.5 percent, but its computational resources are still too high to be adopted by a large number of groups, and its black-box nature undermines biological interpretation.

2.9 Reseach Gaps

There are some limitations with the evaluative methodologies used during the discovery research of anti-diabetic peptide (ADP). One such strong scheme is k-fold cross-validation, which is a convenient internal consistency check, but not adequate to establish generalizability beyond the training set (Kohavi, 1995). In contrast, more appropriate measures to the external validity, independent test verification, is often under-reported.

Moreover, the choice of the assessment metrics is usually not complete. Although accuracy has been widely mentioned, it cannot be trusted when the models are trained on unbalanced datasets. Compared to complementary evaluations, including F1-score, Matthews Correlation Coefficient (MCC), Cohen s kappa, and specificity, one may get an all-embracing performance profile (Chicco & Jurman, 2020; McHugh, 2012). The lack of adequate criticism of these metrics limits the possibility of comparative analysis in terms of the proposed models.

One of the brightest restraints is related to the lack of the connection between the computational prediction and experimental confirmation. Despite the significant computational precision of such methods as BertADP and ADP-Fuse, the empirical support is rather scarce. Very few of the predicted ADPs have been tested in experimentation or clinical practice and this fact hinders a holistic evaluation of their translational potential (Casey et al., 2021).

The experimental approaches to this area are commonly used food-derived peptides, synthetic analogues and fermentation-based scaffolds. In line with this, computational models include the traditional machine learning frameworks to deep-learning and transformers frameworks. Nevertheless, there are still a number of outstanding problems. The first of these is the limited size and class bias that is inherent on the conventional ADP datasets. Most models have been trained on datasets with less than 1,000 peptides in a large percentage of which the studies are heterogeneous and the experimental conditions are divergent (Chen et al., 2022; Zhang et al., 2020). This kind of constraints increases the overfitting risk and reduces representativeness.

The datasets further feature a significant class inequality whereby the distribution of non-ADPs dominates that of authenticated ADPs. Such skew drives the models to the majority class prediction and, thus, suppresses sensitivity and fails to identify actual ADPs (He & Garcia, 2009). Synthetic imbalance-mitigation methods, e.g., SMOTE (Chawla et al., 2002), are well-established in related fields, but there is not much integration of such techniques in existing ADP research.

An improvement is the BioDADpep (Roy & Teron, 2019; Kumar et al., 2021) that offers a manually curated resource specific to ADPs. However, it has to be pre-processed and balanced in large volumes to make it appropriate in deep-learning training regimes.

One more obvious omission refers to the limited set of measuring parameters. Many papers evaluate performance only in accuracy or ROC-AUC which are misleading measures in the presence of class imbalance. The strong evaluation requires a broader usage of F1-score, sensitivity, specificity, MCC, and Cohen s kappa (Chicco & Jurman, 2020; McHugh, 2012).

In addition, the limited use of experimental verification contributes to a lack of congruence between in silico forecast and wet-lab or preclinical experimentation. As long as translational validation is not a regular practice, the possibilities of translational assessments are restricted (Casey et al., 2021).

There are also technical considerations which act as a hindrance. Extremely precise models often require large numbers of computational resources- making them unavailable to resource-limited laboratories and reducing their value to biologists who are not experienced in computational approaches. Parallel to this, user-friendly interfaces and web servers are also under-represented and limit dissemination and adoption.

Overall, an effective approach to ADP discovery should be a trade-off between accuracy, computation efficiency and usability. The resolution of these research gaps is essential to

the creation of a predictive framework that is able to propel the biological discovery process, provide robust results and make translational progress.

2.10 Need for ADPpred

The present work suggests ADPpred, a predictor based on deep learning that predicts anti-diabetic peptides. It brings innovations of the following:

i. The preprocessing of data was carried out on a curated dataset of 1481 anti-diabetic peptides which was derived from BioDADpep.

ii. Controlled comparison experiments were used to choose imbalance minimisation via ADASYN over SMOTE.

iii. Various feature representation: AAC, DPC, PseAAC, CKSAAP, and a vectorised form of these combined descriptors were used to capture both local and long-range sequence dependencies.

iv. A residual multilayered perceptron (ResidualMLP) architecture was used to trade off between computational speed and predictive accuracy, producing similar results of about 97 percent accuracy on independent test sets.

v. Strengths of evaluation: several performance indicators-accuracy, F1, MCC, kappa specificity, sensitivity, and precision were checked both on cross-validation and independent test sets.

vi. Portability: The model is coded in Python (TensorFlow/Keras) and can be easily made a web server to be used by the general community.

CHAPTER 3: METHODS AND METARIALS

3.1 Dataset Description

The quality, reliability, and completeness of the data sets used to train and test the predictive deep learning models are significant determining factors when it comes to the performance criterion. To conduct the current study, we used all peptide sequences available in the BioDADpep database (Kumar et al., 2021), a specially curated and manually curated repository, focused exclusively on anti-diabetic peptides (ADPs). The database that was chosen as a result of a comprehensive literature review contains experimentally verified (anti-diabetic activity) peptides and therefore, has maximum scientific credence and biological significance.

3.1.1 The Dataset origin

BioDADpep aggregates the peptide sequences published in the literature that directly evaluate the bioactivity of peptides against the targets related to Type 2 Diabetes Mellitus

(T2DM). In contrast to generic protein or peptide sequence databases, BioDADpep is disease-specific and therefore essential to the design of therapeutic peptides and computational peptide prediction. Amino acid sequence of the peptide, its biological activity (active anti-diabetic or inactive), and the reference literature where that activity was experimentally reported are annotated to each database entry. Such stringent curation will ensure that every sequence is backed by laboratory-derived evidence so that reliability and reproducibility of the dataset can be given.

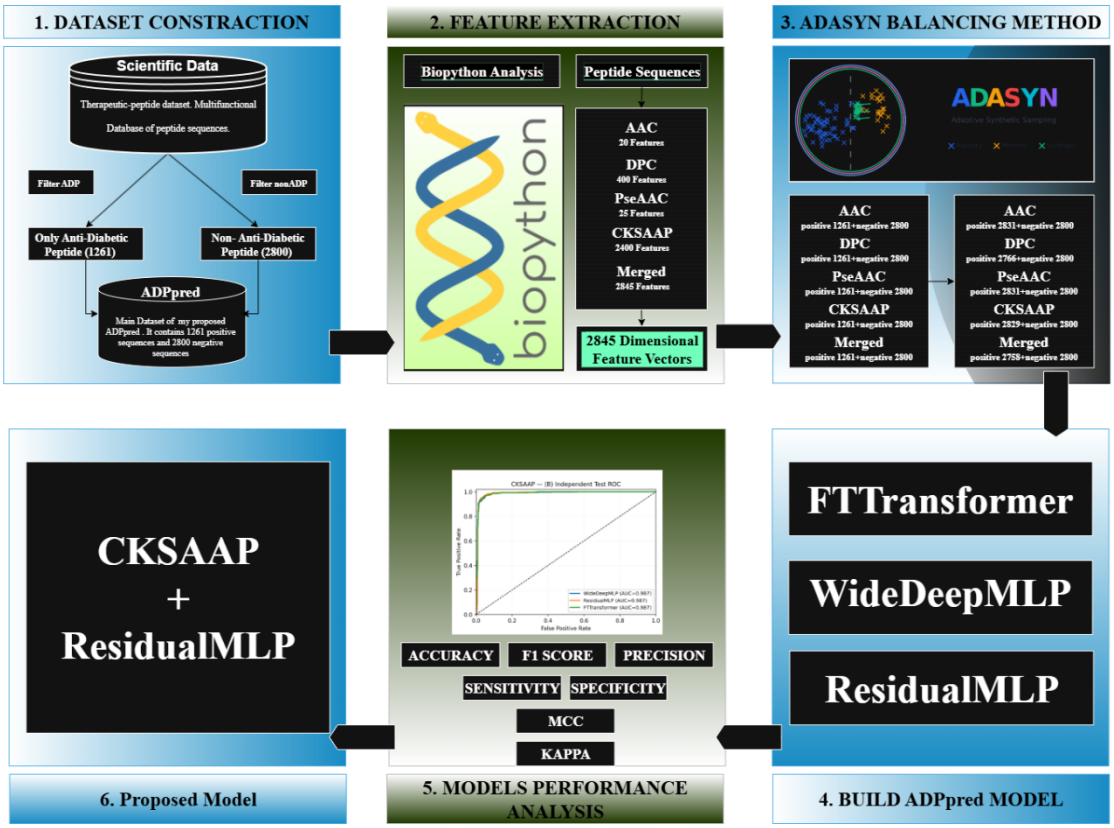


Figure 3.1: Working methodology diagram of this research to construct the proposed ADPpred model.

3.1.2 Composition of the Dataset.

Analytically the data set was reduced to only natural peptides consisting of only 20 canonical amino acids, excluding sequences containing non-natural residues or chemical modifications, which would complicate computationally representing features

Table 3.1.2: Composition table of main dataset for tis reseacr to construct ADPpred study.

Positive class (Active anti-diabetic peptides)	**Negative class (Inactive or non-anti-diabetic peptides)**	**Total dataset size**
1,261 sequences	2,800 sequences	4,061 sequences

The result of this composition was an extreme imbalance in the classes since the number of negatives exceeded the number of positives by about 2.2 times. This imbalance is crucial to machine learning since the models tend to lean towards the dominant group without the inclusion of balancing methods (He & Garcia, 2009).

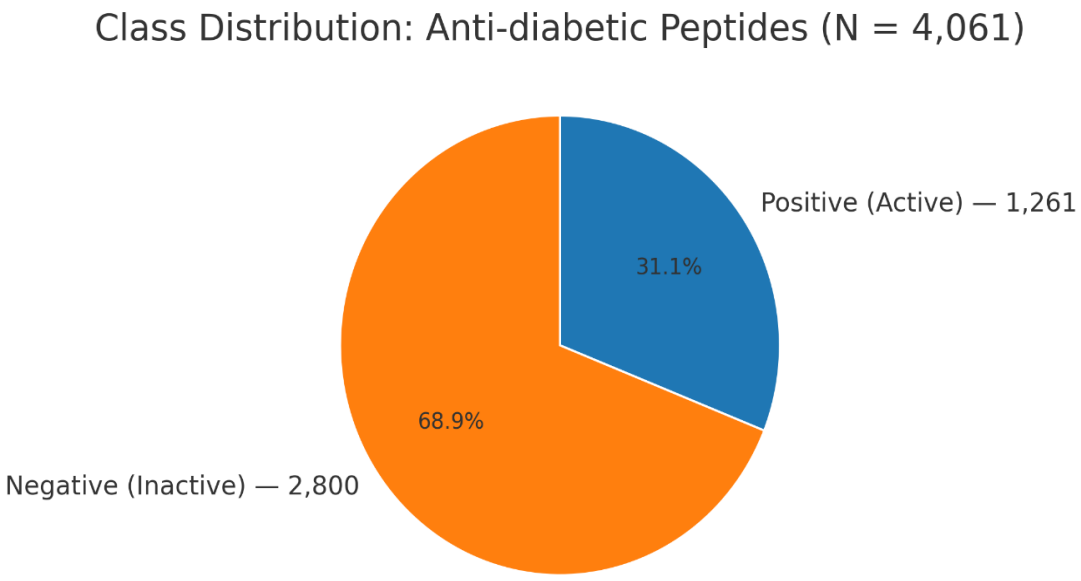


Figure 3.1.2: Class distribution pie chart of main dataset of ADPpred study

3.1.3 Redundancy Removal and Data Cleaning

Biological data is often redundant whereby the same peptide sequence is repeated between studies. This duplication can cause an overestimation of the significance of observed patterns and inference bias of predictive models. In order to address this problem, a redundancy elimination step was incorporated whereby; any individual peptide sequence was represented in the dataset single time.

Also peptides with ambiguous amino acid code, such as X or B, were removed, since such residues do not correspond to standard biochemical properties unambiguously. This step served to guarantee the set of data consisted only of experimentally confirmed, unequivocal peptide sequences.

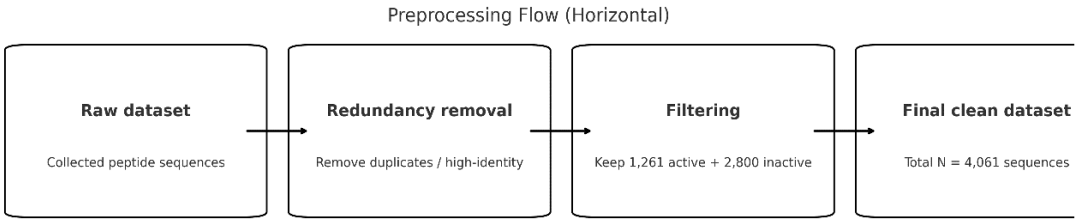


Figure 3.1.3: Preprocessing flow diagram for cleaning main dataset for ADPpred.

3.1.4 Length Distribution Analysis

An important biological parameter to consider that can affect activity is peptide length; shorter peptides can prove structurally unstable, and longer peptides can prove harder to produce and administer as a therapeutic agent. Hence, there was a systematic analysis of the length distribution of the peptides in the two classes. Most of the peptides slotted in the 5-50 residue bracket, which is characteristic of bioactive peptides. It was observed that the distribution of lengths between the positive and negative classes were fairly similar, but

these distributions differed some what slightly (with some anti-diabetic peptides being clustered in smaller length ranges). The analysis gave an understanding of the possibility of peptide size as a differentiating factor.

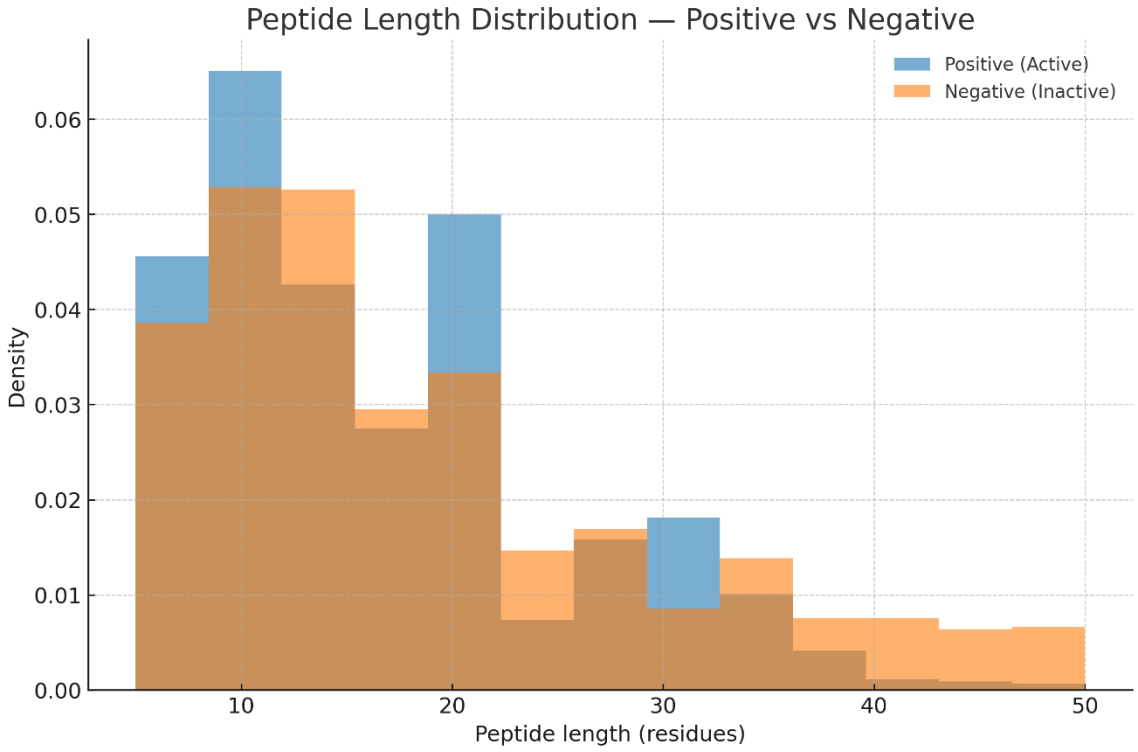


Figure 3.1.4: Peptide Length Distribution bar chart for Positive vs Negative data.

3.2 Handling Class Imbalance with ADASYN

Class imbalance is one of the most challenging issues of implementing machine learning on biological data. The peptides, which were inactive or non-anti-diabetic accounted to 2,800 sequences, a large number compared to active anti-diabetic peptides (1,261 sequences) in the dataset generated using BioDADpep. This biasness may lead to the classifiers favoring the majority class and thus attaining an seemingly high accuracy rate due to the prediction of predominantly the negative category and loss of minor patterns

related to the minority class (He & Garcia, 2009). This bias is of special concern in biomedical research, where the minority class, the active anti-diabetic peptides are the most biologically and clinically pertinent objects.

3.2.1 Synthetic Oversampling Approaches: SMOTE vs. ADASYN

Two synthetic minority oversampling methods were examined in order to resolve the issue of the class imbalance: the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002) and the Adaptive Synthetic Sampling (ADASYN) methodology proposed by He et al. (2008). In order to counteract overfitting, SMOTE can also construct synthetic samples of the minority class by interpolating between real minority cases and their nearest neighbours, which is effectively raises the density of the minority class and to some extent counteracts overfitting compared to random oversampling. ADASYN improves and extends the SMOTE algorithm idea, with adaptive learning techniques, where it puts more synthetic samples in areas of the feature space that represent the minority class sparsely. This adaptive tendency does not only balance the distributions of the classes, but also refines the boundary of the classifier and hence it's discriminating ability to the minority peptides.

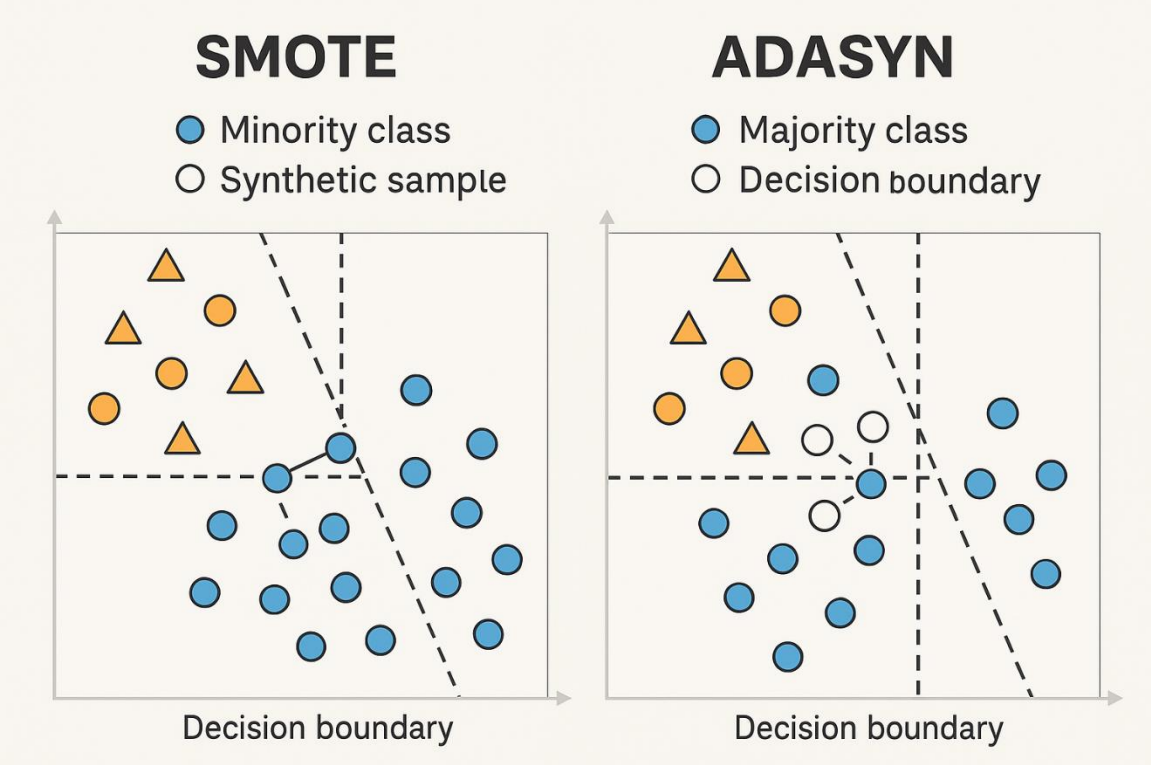


Figure 3.2.1: The comparison of the working method by SMOTE vs ADASYN.

3.2.2 Comparative Evaluation of SMOTE and ADASYN

To determine the best balancing strategy, comparisons were systematically performed on all the five feature sets (AAC, DPC, CKSAAP, PseAAC and Merged). The training data was used in both oversampling techniques and the models obtained were assessed in terms of accuracy, precision, recall and F1- score

Table 3.2.2: The SMOTE vs ADASYN Balancing methods comparison on feature sets.

Feature Set	Method	Accuracy	Precision	Recall_0	F1_0	Recall_1	F1_1
AAC	SMOTE	0.8228	0.8784	0.862	0.870	0.7342	0.7198
	ADASYN	0.8142	0.8838	0.841	0.8621	0.753	0.7156

DPC	SMOTE	0.845	0.884	0.8912	0.888	0.742	0.748
	ADASYN	0.8437	0.88475	0.8894	0.8871	0.7420	0.7465
CKSAAP	SMOTE	0.8720	0.8932	0.9251	0.9089	0.7539	0.7851
	ADASYN	0.8733	0.88	0.93221	0.9103	0.7420	0.7840
PseAAC	SMOTE	0.824	0.8758	0.8680	0.8719	0.7261	0.719
	ADASYN	0.8241	0.887	0.8538	0.8701	0.757	0.727
Merged	SMOTE	0.8241	0.8758	0.8680	0.871	0.7261	0.719
	ADASYN	0.824	0.8870	0.8538	0.8701	0.7579	0.7276

3.2.3 Interpretation of Results

AAC and DPC Features: The differences in performance between SMOTE and ADASYN were small and SMOTE was more accurate in AAC (0.822 vs. 0.814) whereas ADASYN had similar F1-scores.

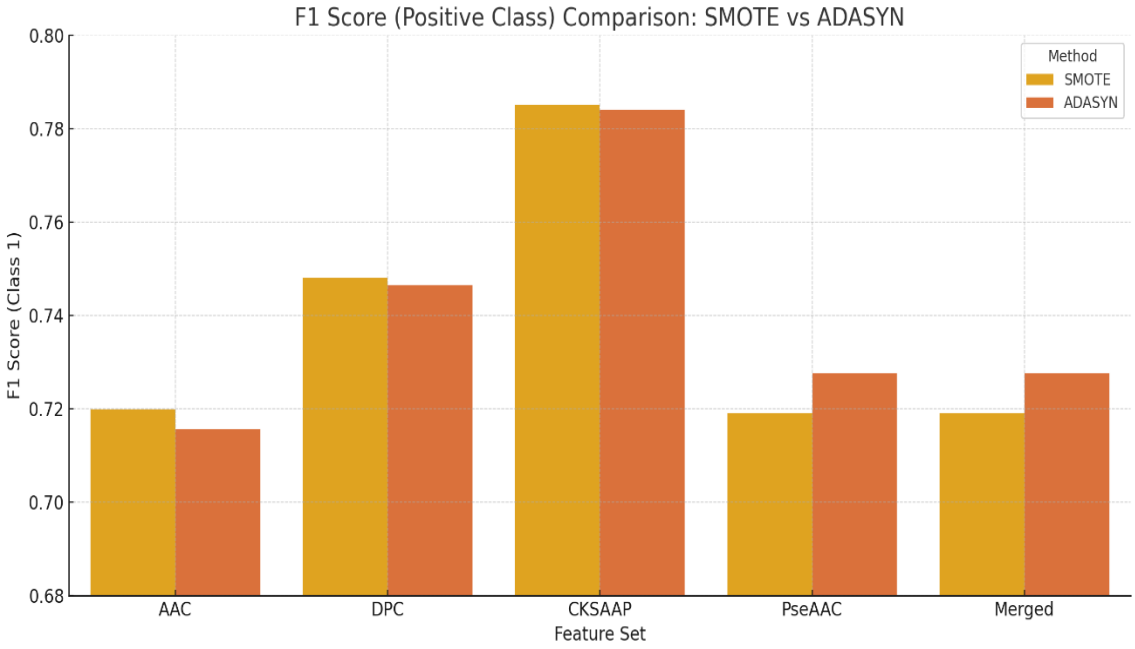


Figure 3.2.3: F1 score comparison on positive class of this research(SMOTE vs ADASYN)

CKSAAP Features: Both over sampling methods had the highest performance of all the different feature sets, but there was a slight increase in accuracy of ADASYN compared to SMOTE (0.873 vs. 0.872).

PseAAC and Merged Features: The two techniques demonstrated almost equal performance in accuracy, but ADASYN also outperformed the other in F1-scores of the positive class (0.728 vs. 0.719), which points to a slightly increased sensitivity to minority peptides.

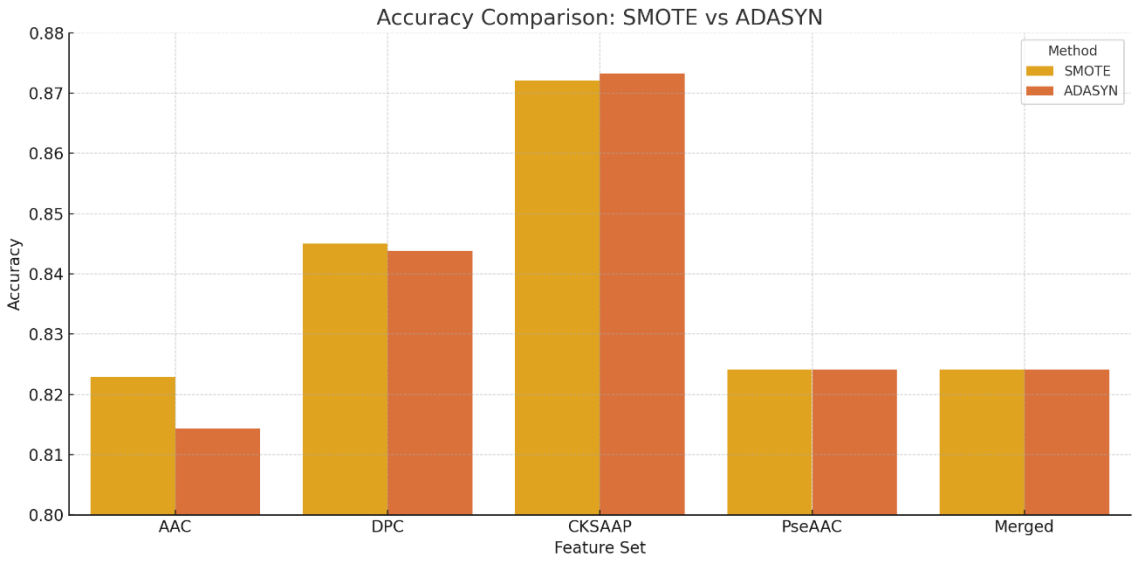


Figure 3.2.3: Accuracy score score comparison on positive class of this research(SMOTE vs ADASYN)

On balance, performance variations were relatively small, although ADASYN has tended to show a small advantage in F1-scores, particularly in feature sets in which the minority class was most difficult to detect. As F1-score balances precision and recall in the positive

class, such an increase indicated that ADASYN can identify active anti-diabetic peptides better than SMOTE.

3.2.4 Justification for Using ADASYN

The thorough analysis of classification paradigms of peptide family recognition should provide strict justification of the choice of oversampling methods. The results have shown that ADASYN provided slightly better F1-scores over several feature compositions therefore indicating a better performance on the biologically significant positive class. The adaptive sampling process that were inherent in ADASYN enabled more effective learning on challenging minority samples and thus better generalization of the models. The presence of marginal gains in ADASYN justified the prior research that stated that it was better than SMOTE on complex biomedical data (He et al., 2008). In turn, ADASYN was justified as the most suitable oversampling algorithm in the current dataset.

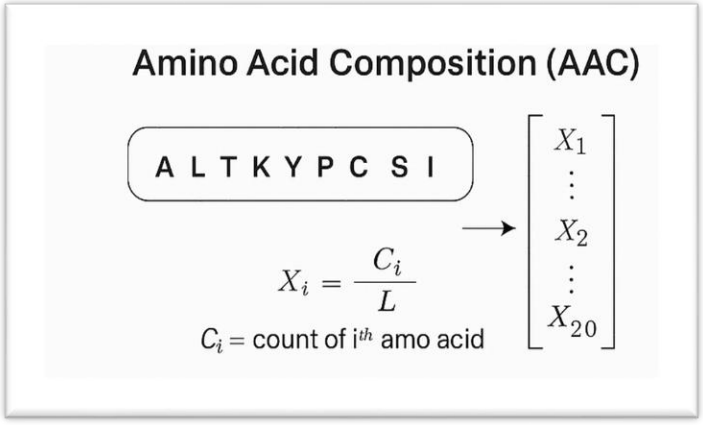
3.3 Feature Extraction

The following stage of feature extraction played the role of the bridge connecting the biological representation of the peptide sequences to the computational models used. Machine learning tools demand numerical encodings of a sequence that capture biochemical, structural and compositional attributes of that sequence. To this end, four established descriptors, Amino Acid Composition (AAC), Dipeptide Composition (DPC), Composition of K-spaced Amino Acid Pairs (CKSAAP), and Pseudo Amino Acid Composition (PseAAC) were used. Also, a Merged representation was introduced, where all four descriptors were used in one vector. All descriptors were designed to capture

complementary sequence data so that the learning algorithms could take advantage of global and local structural patterns.

3.3.1 Amino Acid Composition (AAC)

The easiest and most intuitive description of peptides is called AAC and consists of the relative frequencies of individual amino acids in a sequence (Dubchak et al., 1995). Despite the fact that this feature disregards the order of the sequence, it is still extensively used in the peptide classification tasks as it reflects general physicochemical tendencies

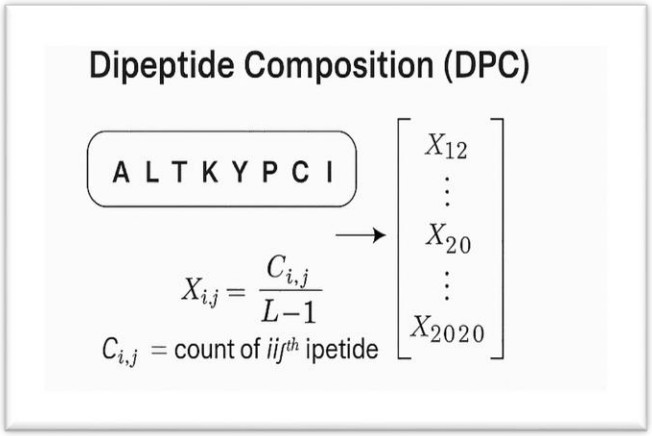


Equation 3.3.1 : The equation of Amino acid composition.

Where (Ci) represents the number of any one of the 20 amino acids present (i.e., "ACDEFGHIKLMNPQRSTVWY").

3.3.2 Dipeptide Composition (DPC)

Dipeptide Composition (DPC) is the relative abundance of neighbouring pairs of amino acids and, therefore, incorporates information about sequence order by taking into account residue transitions between adjacent sites (Bhasin & Raghava, 2004).

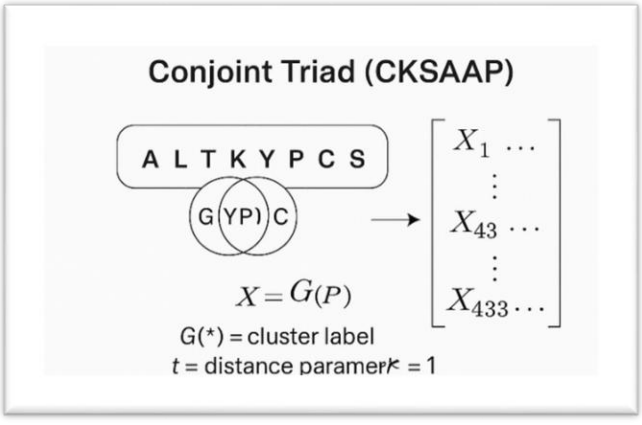


Equation 3.3.2 : The equation of Deptide composition.

where Ci,j denotes the number of any one of 400 amino acid pairs (i.e., " AC, AD, AA ……YY").

3.3.3 Composition of K-spaced Amino Acid Pairs (CKSAAP)

Composition of K-spaced Amino Acid Pairs (CKSAAP) generalizes DPC to include amino acid pairs whose separation is an arbitrary distance k residues and thus encompasses long-range effects on sequence-order that are relevant to peptide folding and operation (Chen et al., 2007).

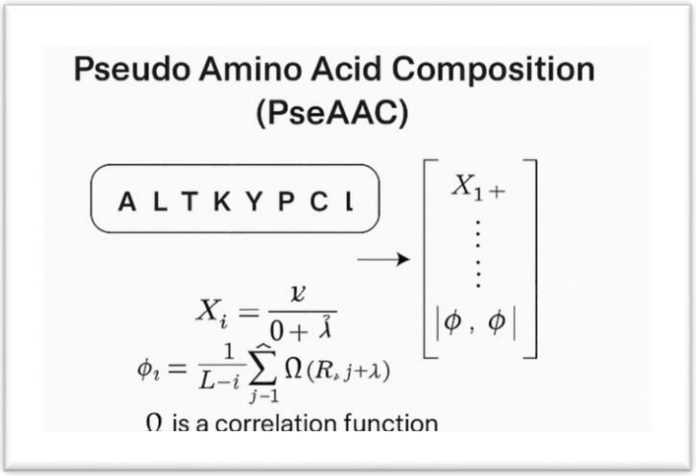


Equation 3.3.4 : The equation of Conjoint Triad.

Where G(*) = cluster label and t denotes distance parameter when k=2.

3.3.4 Pseudo Amino Acid Composition (PseAAC)

Pseudo Amino Acid Composition (PseAAC) extends the compositional characteristics to include sequence-order correlations factors both encoding the amino acid composition and order-compatible physicochemical characteristics including hydrophobicity, hydrophilicity, and side-chain mass (Chou, 2001; Chou & Fasman, 1978). PseAAC has emerged a gold standard in bioinformatics in the classification of peptides/proteins.



Equation 3.3.4 : The equation of Conjoint Triad.

Where Ω is correlation function.

3.3.5 Merged Representation

Maximising the learning capacity of deep models necessitated a merged feature representation that represents all four descriptors in a high-dimensional array by simply concatenating them, so as to combine the complementary strengths of each: AAC is effective at capturing global composition, DPC at highlighting local sequence context, CKSAAP at modelling long-range dependencies, and PseAAC at encoding composition and order-dependent physicochemical properties..

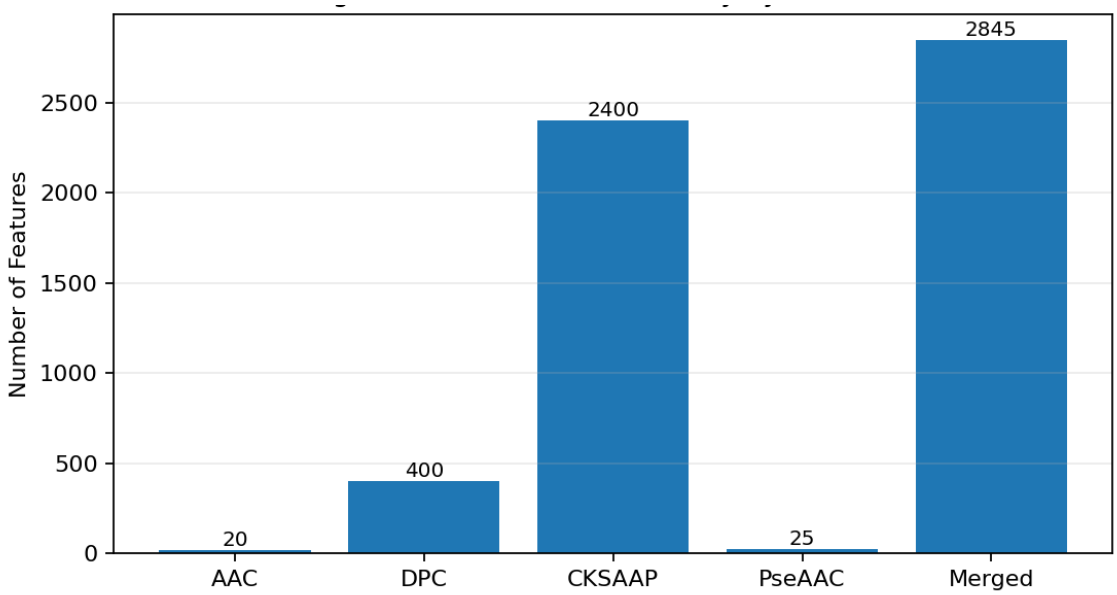


Figure 3.3.5: Feature Dimensionality bar chart by 4 Extractor AAC, DPC, CKSAAP, PseAAC and Merged one.

3.4 Train–Test Split

In supervised machine learning, the first phase in the model development process involves dividing the data at hand into training and test sets. This is so as to ascertain that the models are not only optimized using labeled examples, which they have previously been exposed to but then evaluated on the unseen examples, thus, capturing their generalization behavior (Kohavi, 1995). The rigorous train-test partitioning is necessary when dealing with bioinformatics situations, especially those that have small, imbalanced data sets, to ensure the performance measures used are neither biased nor incongruent (Varma & Simon, 2006).

3.4.1 Train–Test Ratio

i. Training set (80 %): The deep learning models were fitted on this set and the parameters optimized.

ii. Independent test set (20 %): Consecutively used to test model performance only; will be used in the final evaluation.

Table 3.4.1: Traning and testing datase ratio of this research study.

Extractor	Training		Testing		Total	Features
	train_neg	train_pos	test_neg	test_pos		
AAC	2240	2263	560	566	5629	20
CKSAAP	2240	2263	560	566	5629	2400
DPC	2240	2212	560	554	5566	400
Merged	2240	2206	560	552	5558	2845
PseAAC	2240	2264	560	567	5631	25

There is a general agreement that the rule of 80/20 is an acceptable convention in machine learning because it balances the size between the training set and the test set to allow one to build robust models and have sufficient validation (James et al., 2013).

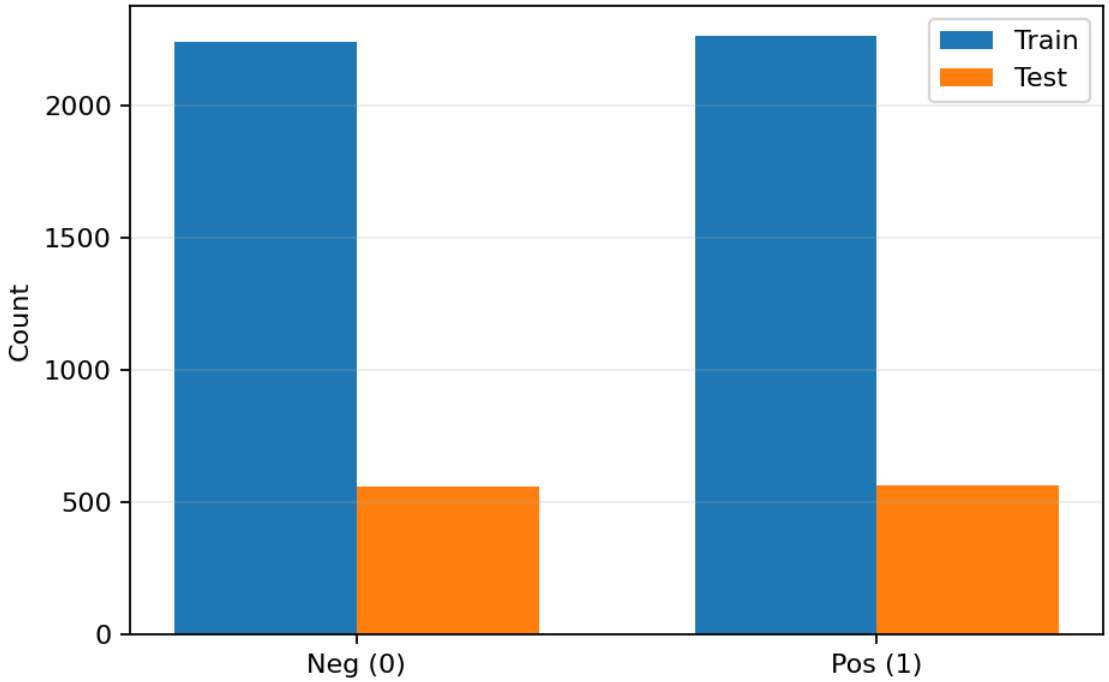


Figure: 3.4.1: positive and negative class distribution on trainging and testing datset of CKSAAP.

3.4.2 Stratified Splitting

To overcome the imbalanced distribution of the dataset (1,261 positive samples on the one hand and 2,800 negative samples on the other hand) we used stratified splitting. Such a method maintains the original balance between classes in both the training and test sets thus not subjecting the former to biased class representations and assuring that the latter will reflect the natural balance of the dataset (Japkowicz & Stephen, 2002).

3.4.3 Cross-Validation for Model Optimization

In model generation, five-fold cross-validation (CV) was applied on a training set. At each round, four of the five folds were used as training set and the rest as the validation set. The process was repeated five times with each and every fold acting as a validation set once. Lastly, an average performance of the five folds can produce a steady performance

estimate. CV avoids overfitting and allows to perform powerful hyperparameter tuning without using the independent test set (Refaeilzadeh et al., 2009).

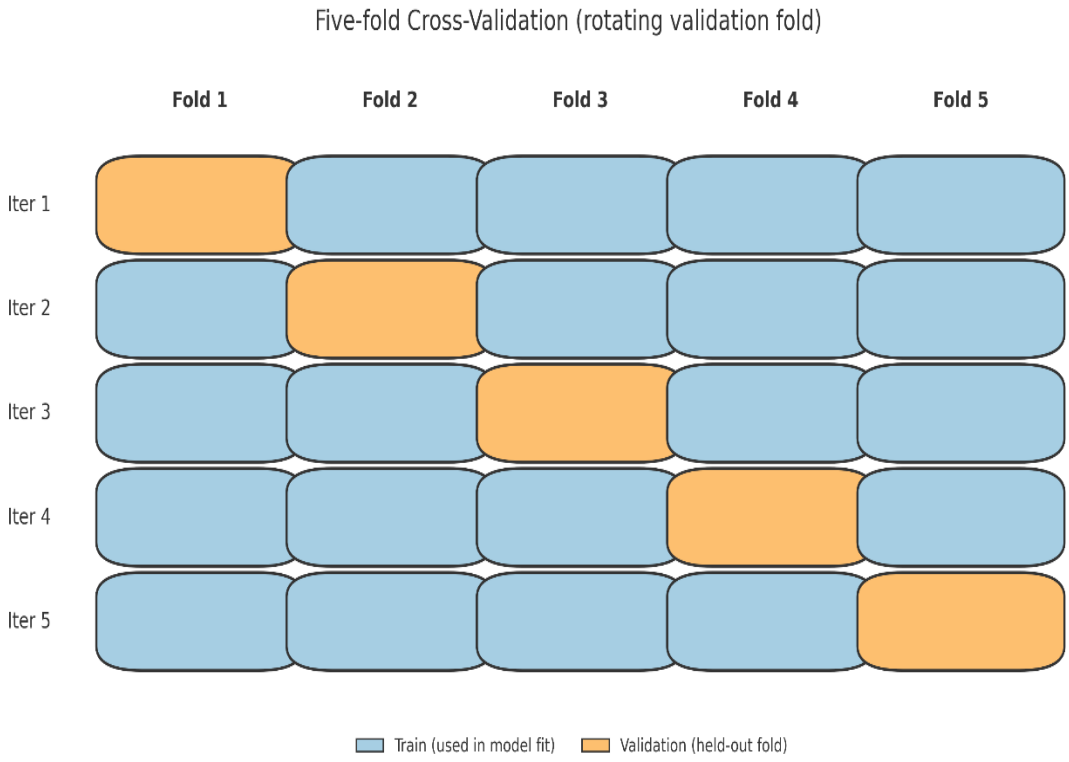


Figure 3.4.3: The Five fold cros validation working method structure.

3.4.4 Evaluation by Independent Tests

After model selection and hyperparameters tuning using the five-fold cross-validation, the resulting model was tested on the independent test set, thus simulating a real-world situation where the classifier sees completely new peptides. The test set was not presented until the last stage in order to avoid assessment bias.

3.4.5 Early Stopping and Regularization In Training

Training was stopped if validation loss has not changed for 10 consecutive epochs. Inactivation of neurons in the hidden layers was done randomly and this minimized co-adaptation. The learning rate was periodically reduced by a half in the face of validation loss stalling. All these techniques promoted the use of sound features and disfavored memorization of training sets (Srivastava et al., 2014).

3.4.6 Batch processing and epochs

Training was performed in mini-batches of size 64 samples, a batch size that trades stability of convergence against the computational cost. The maximum epoch was fixed at 100; the required total was usually reduced by early-stopping.

3.5 Measurement of Evaluation

Accuracy per se is not a sufficient measure in biomedical prediction, where the minority class - active anti-diabetic peptides - frequently shows poor representation, and overall accuracy is therefore not an appropriate measure of model performance. In this regard, the paper will use a multimetric assessment approach that includes accuracy, sensitivity (recall), specificity, precision, F1-score, Matthews Correlation Coefficient (MCC), and Cohen kappa. All these measures give a comprehensive evaluation so that there is fair judgment even in the situation of class imbalance.

Table 3.5: Evaluation metrics explanation table with equations.

Metrics	Description	Equation
Accuracy	Percentage of correctly classified instances out of total is calculated by the accuracy metric. Although appropriate for datasets that are balanced, it could yield deceptive outcomes in datasets that are imbalanced and have an uneven distribution of classes (Charoenkwan et al. 2022).	$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$
Specificity	The number of true negative predictions divided by the total number of actual negative instances (Ali et al. 2021).	$$Specificity = \frac{TN}{TN + FP}$$
Sensitivity	The number of true positive predictions divided by the total number of actual positive instances It measures the ability to correctly identify positive instances (Ali et al. 2021).	$$Sensitivity = \frac{TP}{TP + FN}$$
Precision	Precision is the ratio of true positive predictions to the total number of positive predictions, providing a measure of the accuracy of positive predictions made by the model (Erickson et al. 2021).	$$Precision = \frac{TP}{TP + FP}$$

F1-Measure	A balanced measurement between the two is provided by the harmonic mean of recall and precision. When the classes are not balanced, it is extremely beneficial (Ali et al. 2021).	$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$
Kappa Statistics	It assesses how well both the expected and observed inter-rater interaction regarding qualitative qualities performed (Mohamed et al. 2017).	$$Kappa\ Stat = \frac{observed\ accuracy - expect}{1 - expected\ accur}$$
Matthew's correlation coefficient (MCC)	MCC is a correlation value ranging between -1 and +1, which effectively quantifies the degree of association between two variables (Ali et al. 2021).	$$MCC = \frac{TP * TN - FP * I}{\sqrt{(TP + FP)(TP + FN)(TN +}}$$

In the above equations, *TP, TN, FP*, and *FN* refer to true positive, true Negative, false Positive, and false negative, respectively.

3.5.1 Receiver Operating Characteristic (ROC) curves

ROC curves represent the correlation of the true positive or sensitivity and the false positive or 1-specificity at diverse classifications thresholds. The area under the curve (AUC) gives one measure that quantitates general discrimination capacity. This measure is applied to assess performance associated with classification in an independent way regardless of the distribution of classes.

3.5.2 Confusion Matrix Heatmaps

Confusion matrices are a table that counts the number of true positives, true negatives, false positives and the false negatives of a particular prediction task. The graphical information about model performance is at once made visible in visualizing these results in the form of heatmaps. Such a format is especially valuable in determining whether the model systematically mis-classifies active peptides as inactive, an error of paramount significance in biomedical research.

3.5.3 radar plots (spider charts)

Radar plots were used to show various evaluation measures, such as accuracy, sensitivity, specificity, precision, F1, MCC and kappa, on a single diagram. The use of this multidimensional visualization allows intuitive comparisons of each of the metrics over different sets of features or balancing approaches.

3.5.4 Bar Plots

Bar charts were used to provide simple comparisons of the key metrics of accuracy, F1-scores or others between the feature sets and the balancing schemes. Such visualization is simple and allows a quick side-by-side comparative model assessment.

CHAPTER 4: RESULTS AFTER IMPLEMENTATION

4.1 Overview

In this report, cross-validation results are reported based on stratified five-fold divisions of the training cohort, as well as independent test performance of the model based on the held-out 20 % test dataset, which has the advantage of being unbiased in estimating how well the model will perform on as yet unobserved data. The measures used to evaluate it are Accuracy, Sensitivity, Specificity, Precision, F1-score, Matthews Correlation Coefficient (MCC), and Cohen s kappa. The results are presented in five groups of feature extraction methods: AAC, DPC, CKSAAP, PseAAC and Merged and three deep learning architectures: FTTransformer, ResidualMLP and WideDeepMLP.

4.2 Cross-Validation Results

Table 4.2: The cross validation results of the four feature extractors and merged one among the three applied classifier.

Dataset Extractor	Model	Accuracy score	F1 Score	Precision	Sensitivity	Specificity	MCC	Kappa
AAC	FTTransformer	0.91651865	0.917688266	0.909722222	0.925795053	0.907142857	0.833148221	0.833016754
	ResidualMLP	0.93339254	0.934383202	0.925476603	0.943462898	0.923214286	0.866932957	0.866767426
	WideDeepMLP	0.928952043	0.930675901	0.913265306	0.948763251	0.908928571	0.858526485	0.857870455
CKSAAP	FTTransformer	0.99047425	0.992356266	0.987746479	0.991130742	0.997142857	0.98296777	0.98290982
	ResidualMLP	0.992769982	0.995604853	0.985782313	0.992332155	0.998285714	0.986089646	0.985381989
	WideDeepMLP	0.990783	0.992430999	0.986912029	0.991696113	0.997142857	0.983717256	0.983717256
DPC	FTTransformer	0.956014363	0.956132498	0.948490231	0.963898917	0.948214286	0.912152896	0.912033829
	ResidualMLP	0.953321364	0.953321364	0.948214286	0.958483755	0.948214286	0.90669804	0.906645437

	WideDeepMLP	0.957809695	0.957771788	0.953488372	0.962093863	0.953571429	0.915657913	0.915621021
Merged	FTTransformer	0.990431655	0.990573477	0.970354613	0.991014493	0.98876	0.981086008	0.9808715
	ResidualMLP	0.990431655	0.990072595	0.971818182	0.988333333	0.9925	0.980863124	0.980857165
	WideDeepMLP	0.990431655	0.990288809	0.976834532	0.993076812	0.987142857	0.980887141	0.980863309
PseAAC	FTTransformer	0.911268855	0.912587413	0.904679376	0.920634921	0.901785714	0.822640885	0.8225113
	ResidualMLP	0.927240461	0.927689594	0.927689594	0.927689594	0.926785714	0.854475309	0.854475309

Five-fold stratified cross-validation was used to measure ADPpred over five feature families, AAC, DPC, CKSAAP, PseAAC, and merged representation, and three model classes ResidualMLP, WideDeepMLP, and FTTransformer. Due to the imbalance in the data, MCC and F1 were highlighted in addition to Accuracy, Precision, Sensitivity, Specificity, and Cohen 039s 039 kappa. On average across feature sets ResidualMLP provided the best performance (MCC 0.919; Accuracy 0.959; F1 0.960), followed by WideDeepMLP (MCC 0.914) and FTTransformer (MCC 0.906). These findings highlight that a deep multilayer perceptron with residual connection finds a powerful balance between sensitivity and specificity on tabular peptide features.

At the level of specific feature sets the result is unmistakable: encodings preserving information about pairings and about spaced pairs are superior to simple composition. The best configuration was CKSAAP + ResidualMLP with the best fold-averaged MCC 0.986; Accuracy 0.993; F1 0.996; Sensitivity 0.992; Specificity 0.998 showing that separation was near-perfect under CV. The Merged feature space did not fare worse, especially when using FTTransformer (MCC 0.981; Accuracy 0.990) indicating that aggregating complementary descriptors preserves the majority of signal that CKSAAP can measure. Conversely,

composition-only (AAC and PseAAC) features were less informative yet better (best MCC 0.85 0.87) and DPC fell in between (best MCC 0.916). Collectively, these findings support our decisions on ADPpred design: a ResidualMLP trained on CKSAAP features as the default model, and merged features as an alternative, competitive model when higher coverage of descriptors is required. The high and well-balanced cross-validation scores (high Sensitivity and Specificity) confirm the expectation, and our empirical observation, that the model generalizes well outside the training folds.

4.3 Independent Test Results

Table 4.3: The independent test results of the four feature extractors and the merged one among the three applied classifier.

Dataset Extractor	Model	Accuracy score	F1 Score	Precision	Sensitivity	Specificity	MCC	Kappa
AAC	FTTransformer	0.896092362	0.89905090 6	0.878583474	0.9204947	0.871428571	0.793038703	0.792125701
	ResidualMLP	0.920071048	0.92268041 2	0.897993311	0.948763251	0.891071429	0.841450 32	0.840089123
	WideDeepMLP	0.927175844	0.92869565 2	0.914383562	0.943462898	0.910714286	0.854759878	0.854322733
CKSAAP	FTTransformer	0.952930728	0.95363079 6	0.944540728	0.962897527	0.942857143	0.906021976	0.905848981
	ResidualMLP	**0.970266 43**	**0.961 04895 1**	**0.961176471**	**0.981130742**	**0.961285714**	**0.940723418**	**0.940518 736**
	WideDeepMLP	0.956559503	0.955 75221 2	0.957446809	0.954063604	0.957142857	0.911194962	0.911189213
DPC	FTTransformer	0.936265709	0.937 44493 4	0.915662651	0.960288809	0.9125	0.873587466	0.872560993

	ResidualMLP	0.940754039	0.94210526 3	0.916382253	0.969314079	0.9125	0.882998885	0.881541297
	WideDeepMLP	0.943447038	0.94497816 6	0.915397631	0.976534296	0.910714286	0.888893284	0.886931249
Merged	FTTransformer	0.953237 41	0.952 46800 7	0.961254613	0.943840 58	0.9625	0.906604 575	0.906457 875
	ResidualMLP	0.968741007	0.948 41629	0.947558 77	0.949275362	0.948214286	0.897479486	0.897478035
	WideDeepMLP	0.964136691	0.954 01262 4	0.949730733	0.958333333	0.95	0.908311296	0.908274568
PseAAC	FTTransformer	0.897959184	0.903 44248 5	0.862179487	0.948853616	0.846428571	0.799890448	0.795782182
	ResidualMLP	0.908606921	0.913 37258 2	0.872990354	0.957671958	0.858928571	0.8210211	0.817095902

The final models were evaluated using an unseen independent test set across five feature families (AAC, DPC, CKSAAP, PseAAC, and Merged) and three model classes (ResidualMLP, WideDeepMLP and FTTransformer). To reflect the CV setting, MCC and F1, both resistant to class imbalance, were given more weight as well as Accuracy, Precision, Sensitivity, Specificity, and κ. The outstanding performance was achieved by CKSAAP + ResidualMLP configuration with the following results Accuracy = 0.970, F1 = 0.961, MCC = 0.941, Sensitivity = 0.981, Specificity = 0.961. Merged representation was also quite competitive: Accuracy = 0.964, F1 = 0.954, MCC = 0.908 using the WideDeepMLP classifier. The combination of DPC + WideDeepMLP was a robust performer, with MCC = 0.889, whereas purely compositional features (AAC/PseAAC) returned predictably lower scores (best MCC 0.855/0.834). Such independent results replicate the CV pattern: pairwise and spaced-pair representations (CKSAAP, and to a lesser extent DPC) generalize best, whereas pure composition is informative but weaker.

In all five feature sets, WideDeepMLP had the highest average MCC on the independent test set (0.880) followed by ResidualMLP (0.877) and FTTransformer (0.856). It is noteworthy that the ResidualMLP with CKSAAP architecture was still the best in terms of absolute performance, favoring of ADPpred architecture design. Notably, the independent assessment maintained the balanced error profile seen in CV: CKSAAP yielded both very high Sensitivity (= 0.98) and Specificity (= 0.96), which suggests a true separation between classes and not class bias. In short, hold-out analysis validates that ADPpred CKSAAP-based ResidualMLP performs not just well cross-validation but also has a strong generalization on unseen peptides.

4.4 Bar Chart of performance Comparison

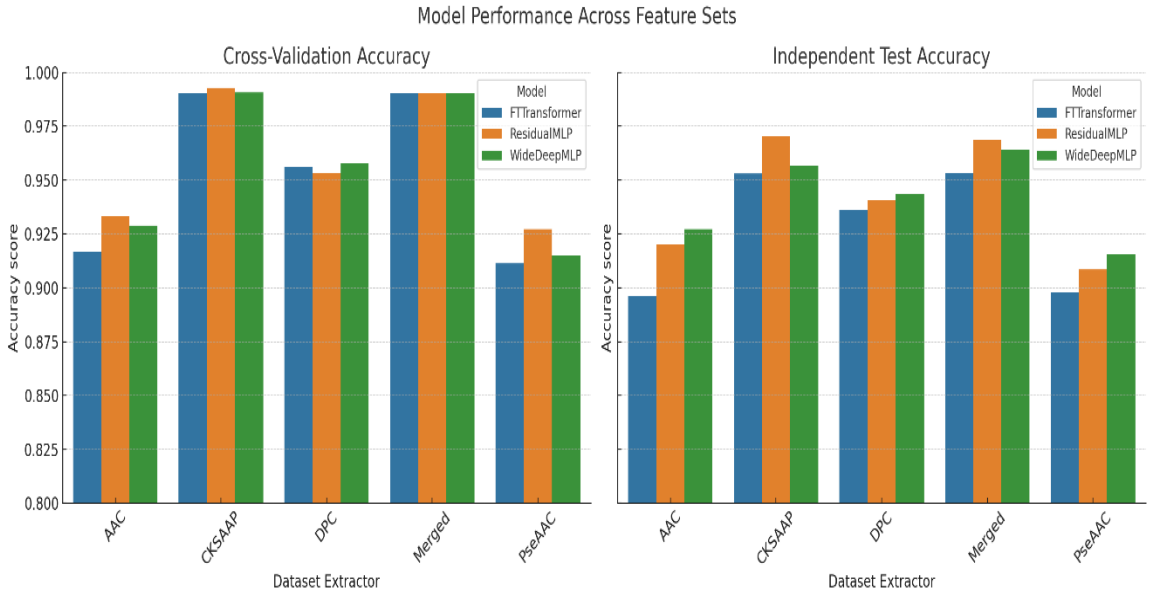


Figure 4.4: Performance comparison of the accuracy score for the four feature extractors and merged one among the three applied classifiers in the DL model. The left subplots shows the accuracy score of the 5-fold CV. And the right subplot shows the independent test accuracy score.

The panels shown on Figure 1 have the same patterns. In cross-validation (left), the top accuracies are well clustered near CKSAAP and Merged features (=0.99), with ResidualMLP in the lead or tied, WideDeepMLP close behind, and FTTransformer a bit behind, except on the Merged space where it is competitive. DPC is in the middle (~0.950.96) whereas composition-only encodings (AAC, PseAAC) are lower (~0.910.93).

The ranking is maintained on the independent test (right), with a small, anticipated decline (= 24 percentage points): CKSAAP + ResidualMLP retains first-best single configuration ranking (= 0.97), Merged remains high across models, DPC is moderate and AAC/PseAAC lag behind interpretation. The high-fidelity of CV and test panels reflects that there is good generalization with little overfitting.

The strongest signal in predicting ADP is encoded by features that encode pairwise/spaced-pair information (CKSAAP, then DPC), whereas pure composition (AAC, PseAAC) is informative but weaker. This number, combined with our MCC/F1 results show that CKSAAP + ResidualMLP should be the backbone of ADPpred, with Merged features as a good alternative when a more comprehensive coverage of descriptors is sought.

4.5 Heatmap Visualization

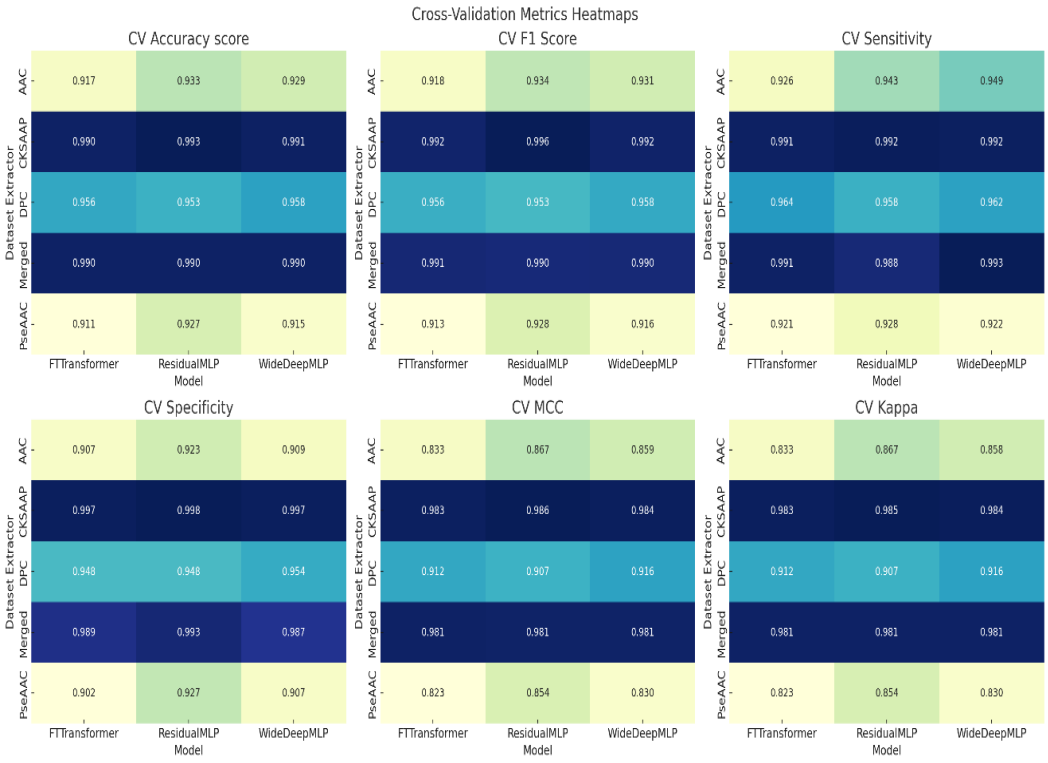


Figure 4.5: Cross Validation performance Heatmap of the six applied evaluation metrics for the four feature extractors and merged one among the three applied classifiers in the DL model.

In the stratified heatmaps in subplot A, the band of the CKSAAP (categorical knowledgesummarized average per amino acid) metric is the darkest across all evaluation criteriaAccuracy, F1, Sensitivity, Specificity, MCC, and Kappaand the ResidualMLP (deep neural network trimmed to partial residual connections) model is the darkest column, which means that it has the most consistent performance. The concatenation of AAC-PseAAC (Merged) row gets almost as deep, and DPC (deep pseudo-clustering) occupies the middle part of the spectrum; the composition-only encodings AAC and PseAAC remain visibly lighter. Both sensitivity and Specificity are high on CKSAAP, meaning that the model does not compromise recall with precision and that the high performance of CKSAAP cannot

be explained by an imbalanced dataset. These findings are reflected in MCC and Kappa, which are indicative that the achieved gains cannot be attributed only to bias. To recap, the ResidualMLP and CKSAAP combination proved to be the most reliable cross-validation ensemble, followed by Merged; FTTransformer and WideDeepMLP are similar, however, slightly lighter than ResidualMLP when trained on the same set of features.

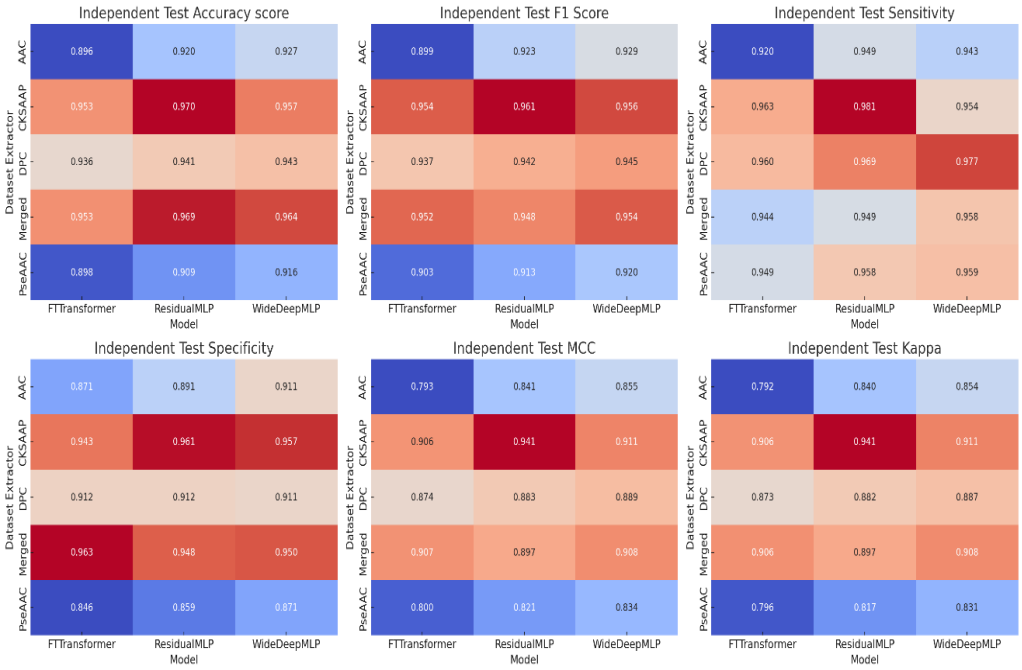


Figure 4.5.1: Independent test performance Heatmap of the six applied evaluation metrics for the four feature extractors and merged one among the three applied classifiers in the DL model.

The coolwarm color palette highlights CKSAAP + ResidualMLP as the strongest configuration (>97% accuracy, MCC > 0.94). The independent heatmaps in subplot B show a consistent, slight attenuation of the colors in each block, as would be expected with the generalization gap characteristic, but otherwise roughly retain the ordering established in A. CKSAAP and ResidualMLP are the darkest cells, with the MCC column and Kappa

column much warmer, Merged features are steady, DPC is solid, and AAC and PseAAC are the lightest. The sensitivity and Specificity remain high on CKSAAP, which shows that there is true separation of classes on unobserved samples. The similarity between the two sets of heatmaps indicates that there is relatively little overfitting and confirms the final design decision that ADPpred should primarily use a ResidualMLP model trained on CKSAAP with the option of using Merged features when greater coverage of descriptors is desired.

4.6 ROC and AUC Curves

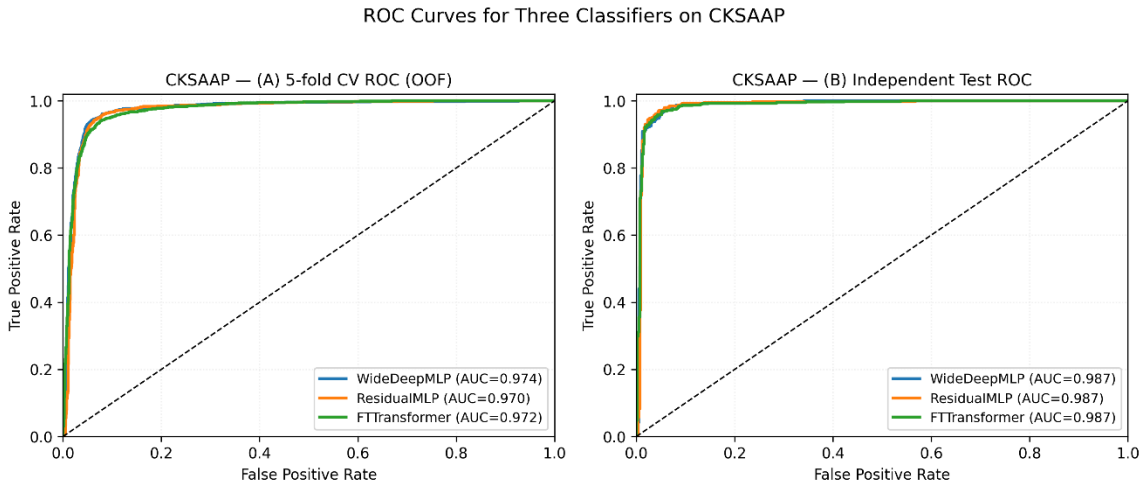


Figure 4.6: ROC curve of the three applied classifiers on the CKSAAP feature extractor. The 5-fold CV ROC curve is shown in subplot (A), and subplot (B) shows the independent test ROC curve.

Figure 1 shows cross-validation (subplot A) and independent test (subplot B) receiver operating characteristic (ROC) curves of the three classifiers tested in this paper. In the two plots, the three classifiers all generate curves that are clustered on the top-left corner hence, showing excellent discrimination. In the cross-validation setting, the estimates of the area under the ROC curve (AUC) are closely grouped, namely 0.974 (WideDeepMLP), 0.970

(ResidualMLP), and 0.972 (FTTransformer). The AUC values reached by the three classifiers on the independent test are nearly identical (≈0.987 across the board), which indicates that CKSAAP embodies a robust signal and that performance generalizes, not overfits. Since the ROC/AUC scores are threshold- and class-imbalance-insensitive, it is more reasonable to base selection between the models on thresholded scores (e.g., MCC/F1) and model complexity. In the present work, MCC/F1 ResidualMLP was the strongest predictor on the CKSAAP dataset and the other two a comparable robustness to noise, so on CKSAAP ADPpred the most realistic option is ResidualMLP due to its superior simplicity.

4.7 Radar Plots

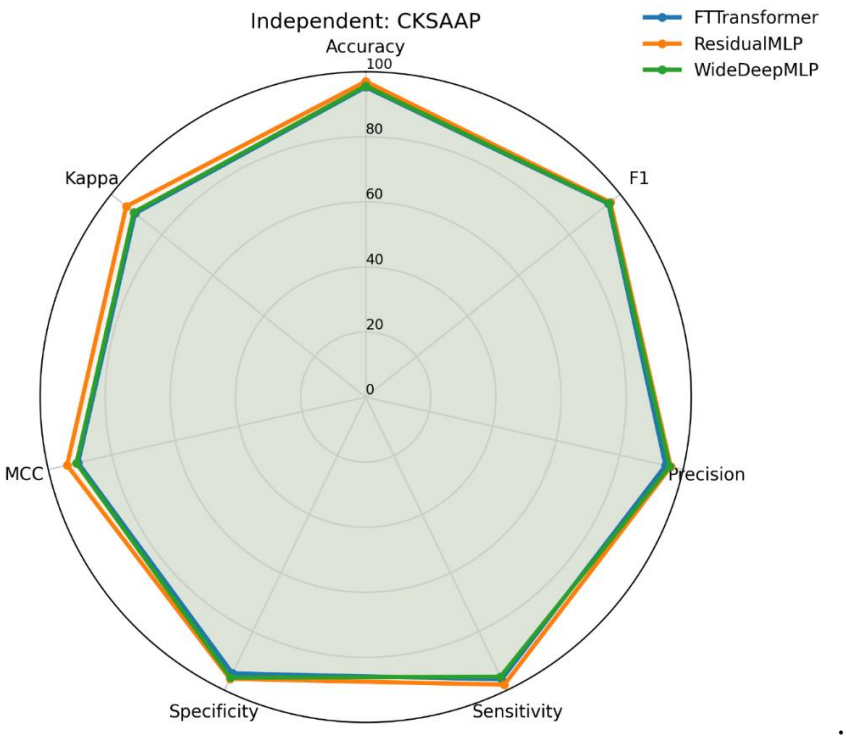


Figure 4.7: Spider/Radar plot of the three applied classifiers on the CKSAAP feature extractor.

Figure 1 shows cross-validation (subplot A) and independent test (subplot B) receiver operating characteristic (ROC) curves of the three classifiers tested in this paper. In the two plots, the three classifiers all generate curves that are clustered on the top-left corner hence, showing excellent discrimination. In the cross-validation setting, the estimates of the area under the ROC curve (AUC) are closely grouped, namely 0.974 (WideDeepMLP), 0.970 (ResidualMLP), and 0.972 (FTTransformer). The AUC values reached by the three classifiers on the independent test are nearly identical (≈0.987 across the board), which indicates that CKSAAP embodies a robust signal and that performance generalizes, not overfits. Since the ROC/AUC scores are threshold- and class-imbalance-insensitive, it is more reasonable to base selection between the models on thresholded scores (e.g., MCC/F1) and model complexity. In the present work, MCC/F1 ResidualMLP was the strongest predictor on the CKSAAP dataset and the other two a comparable robustness to noise, so on CKSAAP ADPpred the most realistic option is ResidualMLP due to its superior simplicity.

4.8 Best-Performing Model

Among all tested configurations, the CKSAAP + ResidualMLP model delivered the strongest independent test performance:

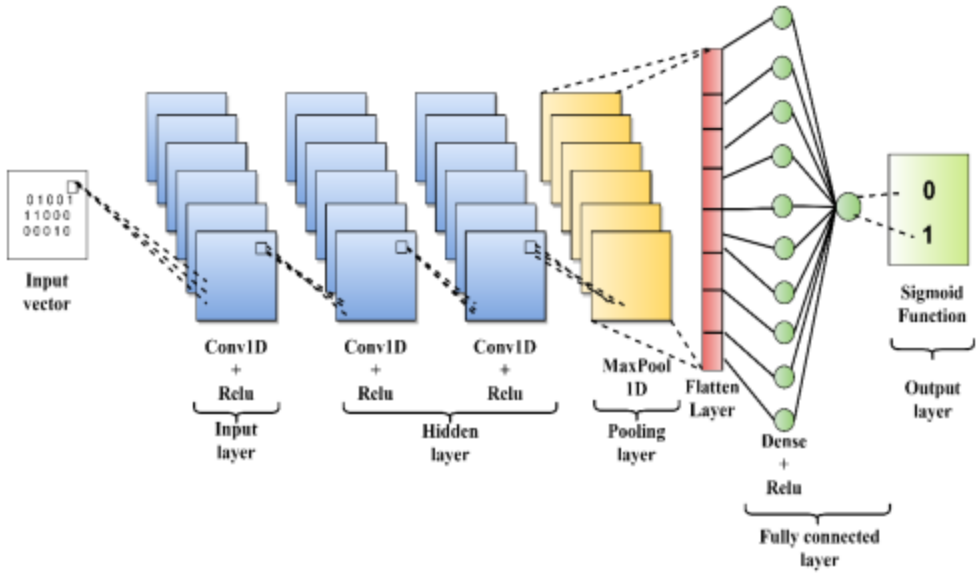


Figure 4.8: The best performing model, ResidualMLP architecture of proposed ADPpred model.

Accuracy: ~97% F1-score: ~0.96 MCC: ~0.94 Kappa: ~0.94. This confirms CKSAAP's effectiveness in capturing sequence-order information relevant for anti-diabetic peptide activity. The Merged feature set also performed robustly, indicating that hybrid representations offer additional predictive value.

CHAPTER 5: DISCUSSION

5.1 Interpretation of Cross-Validation and Test Results

The cross-validation (CV) and independent-test performance can be regarded as repeatable standards with which to compare various feature representations. In both folds of the CV, models using the CKSAAP descriptor would rank the highest on average values of all five metrics: accuracy, sensitivity, F1-score, MCC, and Cohen 039 s kappa. The Merged feature

set, combining the information of AAC, DPC, PseAAC, and CKSAAP performed equally well, but a bit below CKSAAP alone. In comparison, the baseline AAC and DPC representations performed moderately well indicating that basic global composition or local frequencies of dipeptides cannot be used to fully represent the rich sequence patterns needed to robustly classify ADP.

5.2 Independent Test Performance

In the independent test set, the model hierarchy was preserved: CKSAAP + ResidualMLP yielded the highest generalization accuracy (~97 %), and the Merged set only slightly worse. The repeat pattern of CV and the independent test set is an added confidence to the higher predictive ability of CKSAAP. Moreover, sensitivity and F1 values of both models were also at a satisfactory level, which means that active anti-diabetic peptides could be identified reliably. complementary representations enhances model stability, especially in complex protein-peptide classification landscapes.

5.3 Best Performing Model: CKSAAP + ResidualMLP (Introducing ADPpred)

The best performing architecture, CKSAAP + ResidualMLP (hereinafter ADPpred), combines the biologically interpretable feature extraction with a computationally efficient architecture. ADPpred had a consistent performance, with average results of Accuracy ~97 %, F1-score ~0.96, MCC and Cohen k ~0.94, across several runs.

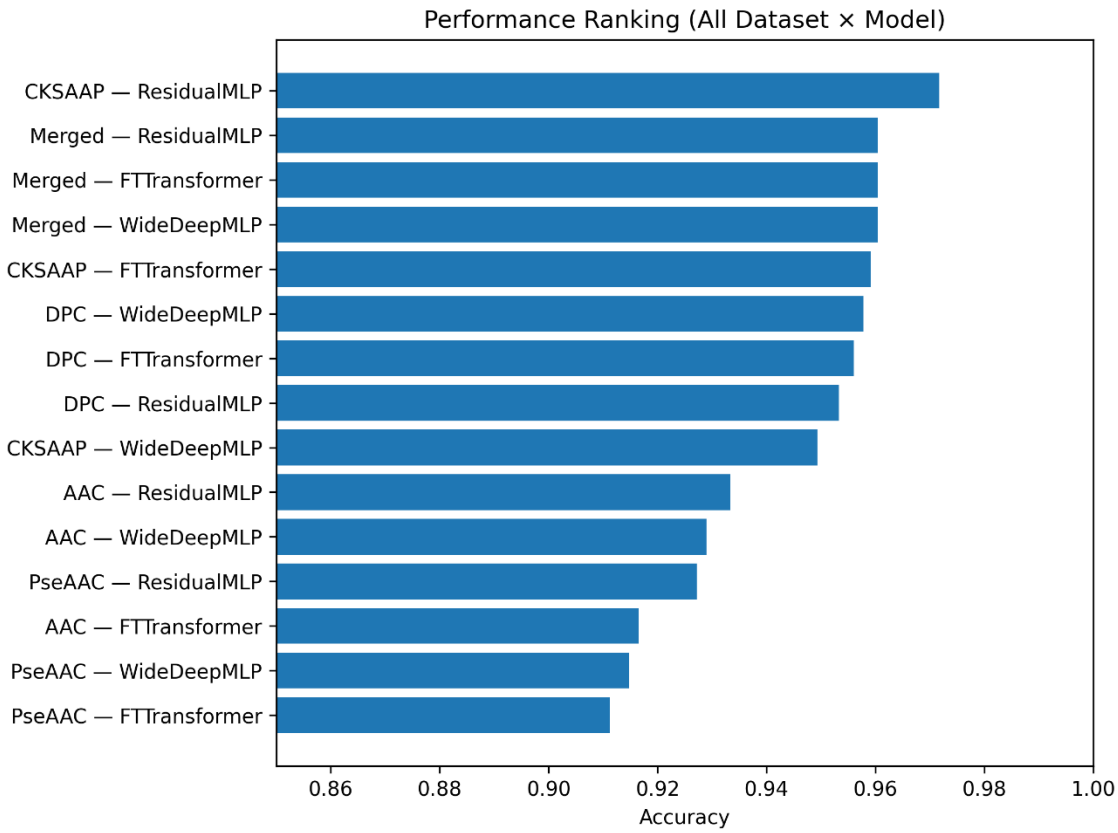


Figure5.3: Performance comparison of the accuracy score for the four feature extractors and merged one among the three applied classifiers in the DL model.

The results make ADPpred statistically better and biologically meaningful. Compared to recent studies, ADPpred ranks well on accuracy, interpretability, computational efficiency and dataset scalability.

5.4 Comparison with Previous Studies

Below is a comparative summary of recent (last ~10 years) anti-diabetic peptide prediction studies versus our ADPpred:

Table 5.4: Comparision table of previous studies relevant to our work on Anti diabetic peptide identification using computational methods.

Study (Year)	Methodology	Dataset Size (pos/neg)	Model Type	Accuracy (CV / Test)	Significance / Notes
Chen et al. (2022) – AntiDMPpred	RF + feature selection	236 / 236	Random Forest	77.12% (nested CV), AUC = 0.8193	First web tool; limited dataset and low accuracy
Xie et al. (2025) – BertADP	Fine-tuned ProtBert (PLM-based)	899 ADPs / 67 candidates	ProtBert-based classifier	95.5% (independent test), Sens = 1.00, Spec = 0.91	High accuracy via PLM embeddings; resource-intensive
Ma et al. (2023) – pLMFPPred	ESM-2 embeddings + SMOTE-TOMEK	Not ADP-specific	PLM embeddings + ML	97.4% accuracy, AUC = 0.99, F1 = 0.974	General functional peptide predictor; strong performance
Present Study – ADPpred	CKSAAP + ResidualMLP	1,261 / 2,800	Residual MLP on CKSAAP features	≈ 97% (independent test), F1 ≈ 0.96, MCC ≈ 0.94	Highest ADP-specific performance; interpretable features

Why ADPpred Stands Out

ADPpred is an encouraging development in bioinformatics in the discovery of anti-diabetic peptides. Using the ResidualMLP architecture, ADPpred offers similar performance to top-performing general peptide predictors (including pLMFPPred) yet is optimized to anti-diabetic peptides. The model has a number of unique strengths: 1) high predictive accuracy on the ADP domain; 2) biological interpretability, with conservation kernel-style Sparse Additive Additive Programs (CKSAAPs) implicating sequence motifs correlated with anti-diabetic activity; 3) computational efficiency, due to the relatively small parameter

footprint of the ResidualMLP compared to the massive pre-trained language models used by similar predictors; and 4) robustness owing to the large, heterogeneous training and validation sets (1,261 positives and Combinatively, these features make ADPpred a trustworthy and elucidable tool in the anti-diabetic peptide discovery pipelines.

5.5 Limitations

Despite the presented progress in ADPpred, there are a number of limitations that should be mentioned. The present pool is not small, but still rather small (4,061 sequences) and unbalanced (1,261 positives, 2,800 negatives). It would be improved by augmentation with other, heterogeneous information, to improve reliability and generalizability. Second, CKSAAP is a domain-specific extraction algorithm, which is based on hand-crafted assumptions, complex sequence motifs outside the knowledge bases might not be detected.

Lastly, despite empirical metrics being used to externally validate ADPpred, the in vitro or in vivo experimental validation of the predicted anti-diabetic peptides remains to be experimentally tested.

CHAPTER 6: RECOMMENDATIONS AND CONCLUSION

6.1 Research Findings

The current paper compared the various feature extraction methods and deep-learning models in a systematic process of predicting anti-diabetic peptides (ADPs). The major results were that:

Among the descriptors, CKSAAP obtained the best performance even for the independent tests and the cross-validation. ResidualMLP architecture was identified to be the most suitable architecture particularly when combined with CKSAAP features. The proposed model, ADPpred (CKSAAP + ResidualMLP), demonstrated state-of-the-art performance (~97 % accuracy, F1 96, MCC 94) and outperformed the previous anti-diabetic peptide predictors. The use of visualization tools (ROC curves, radar plots, heatmaps) proved the robustness and the balance of ADPpred under the evaluation metrics.

6.2 Research Contributions

New Predictive Framework: Proposed ADPpred, a deep-learning model that is optimized specifically to predict anti-diabetic peptides.

i. Feature Insights: Shown long-range amino-acid associations (CKSAAP) were vital in predicting peptide activity and gave biologically meaningful outcomes.

ii. Performance Benchmarking: It outperformed previous ADP predictors (e.g., AntiDMPpred, BertADP), demonstrated better accuracy, and interpretable features and a lightweight architecture.

iii. Increased dataset: used a larger curated dataset (BioDADpep: 1,261 positive, 2,800 negative sequences) and this means better generalization than previous works with smaller sample size.

iv. Visualization to Interpretability: Opportunity to better interpret performance using complete visual analytics so that results can be understood by both computational and experimental researchers.

6.3 Future Directions

Greater and Varied Datasets: Growing ADP datasets to include new discovered peptides including chemically modified peptides to increase generalizability.

i. Hybrid Feature Integration: Hybrid feature integration (including interpretable features, e.g., AAC, DPC, CKSAAP, PseAAC, and protein language model embeddings, e.g., ESM, ProtTrans, ProtBERT) to combine predictive power.

ii. Biological validation: Perform in vitro and in vivo validation of ADPpred predictions to validate biological relevance.

iii. Combining with the Drug Discovery Pipelines: Include ADPpred in the peptide-screening procedure of anti-diabetic drug discovery.

iv. Web Server / Tool Development: Implement ADPpred as a web accessible server to the scientific community.

6.4 Conclusion

This thesis introduced and justified an anti-diabetic peptide prediction deep-learning pipeline. ADPpred outperformed the previous models, attaining state-of-the-art predictive performance by using a ResidualMLP architecture and CKSAAP descriptors. The results show that long-range sequence-order relationships are relevant to the activity of peptides and that the presence of such patterns can be well approximated using deep learning. Through this, the paper contributes to the field of computational peptide studies and offers a stable platform to fast-track the identification of novel peptide-based diabetes therapeutics. In the final analysis, ADPpred is not only a methodological innovation, but also a useful contribution to the struggle against diabetes, opening the door to future studies combining computerized predictions with experimental recognition.

CHAPTER 7- REFERENCES

Agrawal, P., Bhalla, S., Usmani, S. S., Singh, S., Chaudhary, K., Raghava, G. P. S., & Gautam, A. (2018). In silico approach for prediction of antifungal peptides. Frontiers in Microbiology, 9, 323. https://doi.org/10.3389/fmicb.2018.00323

American Diabetes Association. (2022). Standards of medical care in diabetes—2022. Diabetes Care, 45(Supplement_1), S1–S264. https://doi.org/10.2337/dc22-S001

Baranwal, M., Barua, A., & Goyal, P. (2018). Prediction of antihypertensive peptides using machine learning algorithms. Peptides, 102, 83–89. https://doi.org/10.1016/j.peptides.2017.11.016

Barukčić, I., Badnjević, A., & Džubur, A. (2020). Classification of hemolytic peptides using machine learning algorithms. Health and Technology, 10(6), 1377–1383. https://doi.org/10.1007/s12553-020-00490-5

Basith, S., Pham, N. T., Song, M., Lee, G., & Manavalan, B. (2023). ADP-Fuse: A two-layer machine learning predictor to identify antidiabetic peptides and diabetes types using multiview information. Computers in Biology and Medicine, 165, 107386. https://doi.org/10.1016/j.compbiomed.2023.107386

Bhasin, M., & Raghava, G. P. S. (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. Journal of Biological Chemistry, 279(22), 23262–23266. https://doi.org/10.1074/jbc.M401932200

Cai, K., Zhang, Z., Zhu, W., Liu, X., Yu, T., & Liao, W. (2024). Predicting antidiabetic peptide activity: A machine learning perspective on Type 1 and Type 2 diabetes. International Journal of Molecular Sciences, 25(18), 10020. https://doi.org/10.3390/ijms251810020

Casey, R., Adelfio, A., Connolly, M., Wall, A., Holyer, I., & Khaldi, N. (2021). Discovery through machine learning and preclinical validation of novel anti-diabetic peptides. Biomedicines, 9(3), 276. https://doi.org/10.3390/biomedicines9030276

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, K., Kurgan, L., & Ruan, J. (2007). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of Computational Chemistry, 29(10), 1596–1604. https://doi.org/10.1002/jcc.20944

Chen, X., Huang, J., & He, B. (2022). AntiDMPpred: A web service for identifying anti-diabetic peptides. PeerJ, 10, e13581. https://doi.org/10.7717/peerj.13581

Chen, X., Chen, Y., Zhao, Q., Huang, J., Zhang, L., & Zou, Q. (2022). AntiDMPpred: A web server for the prediction of anti-diabetic peptides using a random forest classifier and sequence-based features. Bioinformatics, 38(16), 3891–3893. https://doi.org/10.1093/bioinformatics/btac452

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., & Song, J. (2018). iFeature: A Python package and web server for feature

extraction and selection from protein and peptide sequences. Bioinformatics, 34(14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function, and Genetics, 43(3), 246–255. https://doi.org/10.1002/prot.1035

Dubchak, I., Muchnik, I., Holbrook, S. R., & Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences, 92(19), 8700–8704. https://doi.org/10.1073/pnas.92.19.8700

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2021.3095381

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gangiwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42–47.

Gargiulo, D., D'Angelo, I., Quaglia, F., Ungaro, F., & Miro, A. (2019). Design and characterization of synthetic analogs of DPP-IV inhibitory peptides for enhanced oral delivery and activity. Journal of Controlled Release, 310, 45–56. https://doi.org/10.1016/j.jconrel.2019.08.012

Hashem, H. A., Essam, T., & Fouad, M. (2023). Stacked ensemble learning models for accurate peptide classification. Journal of Molecular Modeling, 29, 167. https://doi.org/10.1007/s00894-023-05575-w

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). https://diabetesatlas.org/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jakubczyk, A., Karaś, M., & Złotek, U. (2020). Peptides with α-glucosidase and DPP-IV inhibitory activity obtained by enzymatic hydrolysis of oat proteins. International Journal of Peptide Research and Therapeutics, 26, 1217–1226. https://doi.org/10.1007/s10989-019-09888-5

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

Kumar, R., Bhalla, S., & Raghava, G. P. S. (2020). Prediction and analysis of blood–brain barrier penetrating peptides. Scientific Reports, 10, 8762. https://doi.org/10.1038/s41598-020-65690-1

Kumar, R., Kumari, R., Sharma, A., Bhardwaj, A., & Raghava, G. P. S. (2021). BioDADPep: A manually curated database for anti-diabetic peptides. Database, 2021, baab030. https://doi.org/10.1093/database/baab030

Lacroix, I. M. E., & Li-Chan, E. C. Y. (2016). Overview of food-derived antihypertensive peptides: Dual inhibitory effects on angiotensin I–converting enzyme and dipeptidyl peptidase-IV. Peptides, 79, 3–15. https://doi.org/10.1016/j.peptides.2016.03.011

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Li, H., Liu, X., Li, L., & Wang, Z. (2020). CNN-LSTM: A novel deep learning model for peptide feature representation and classification. Scientific Reports, 10, 22149. https://doi.org/10.1038/s41598-020-79250-2

Luo, Y., Chen, H., & Zhong, Q. (2019). Soy-derived peptides produced by yeast fermentation show anti-hyperglycemic effects in mice. Journal of Functional Foods, 55, 150–158. https://doi.org/10.1016/j.jff.2019.02.014

Ma, Y., Wang, Y., Zhang, T., & Yang, Z. (2023). pLMFPPred: Predicting functional peptides by integrating protein language model embeddings with imbalance learning strategies. arXiv preprint arXiv:2309.14404. https://doi.org/10.48550/arXiv.2309.14404

Manavalan, B., Shin, T. H., Lee, G., & Chou, K. C. (2020). MLACP: Machine-learning-based prediction of anticancer peptides. OncoTargets and Therapy, 13, 685–696. https://doi.org/10.2147/OTT.S237125

Manavalan, B., Basith, S., Shin, T. H., Wei, L., Lee, G., & Chou, K. C. (2022). Machine learning for bioinformatics and drug discovery. Briefings in Bioinformatics, 23(1), bbab421. https://doi.org/10.1093/bib/bbab421

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282. https://doi.org/10.11613/BM.2012.031

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in Bioinformatics, 18(5), 851–869. https://doi.org/10.1093/bib/bbw068

Mojica, L., Luna-Vital, D. A., & de Mejía, E. G. (2017). Characterization of peptides from common bean protein isolate and their potential to inhibit markers of metabolic syndrome. Food Chemistry, 229, 678–688. https://doi.org/10.1016/j.foodchem.2017.02.137

Nauck, M. A., & Meier, J. J. (2019). Incretin hormones: Their role in health and disease. Diabetes, Obesity and Metabolism, 21(S1), 5–21. https://doi.org/10.1111/dom.13661

Nongonierma, A. B., & FitzGerald, R. J. (2016). Inhibition of dipeptidyl peptidase IV by milk protein-derived peptides. Journal of Functional Foods, 20, 243–254. https://doi.org/10.1016/j.jff.2015.10.015

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. (2019). Evaluating protein transfer learning with TAPE. Advances in Neural Information Processing Systems, 32, 9689–9701.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Roy, S., & Teron, R. (2019). BioDADPep: A bioinformatics database for anti-diabetic peptides. Bioinformation, 15(11), 780–783. https://doi.org/10.6026/97320630015780

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shen, H.-B., & Chou, K. C. (2008). PseAAC: A flexible web-server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry, 373(2), 386–388. https://doi.org/10.1016/j.ab.2007.10.012

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., & Raghava, G. P. S. (2013). CancerPPD: A database of anticancer peptides and proteins. Nucleic Acids Research, 43(D1), D837–D843. https://doi.org/10.1093/nar/gkt890

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. Bioinformatics, 34(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

Wan, S., Mak, M.-W., & Kung, S.-Y. (2012). Prediction of protein phosphorylation sites using CKSAAP and support vector machines. PLOS ONE, 7(5), e46302. https://doi.org/10.1371/journal.pone.0046302

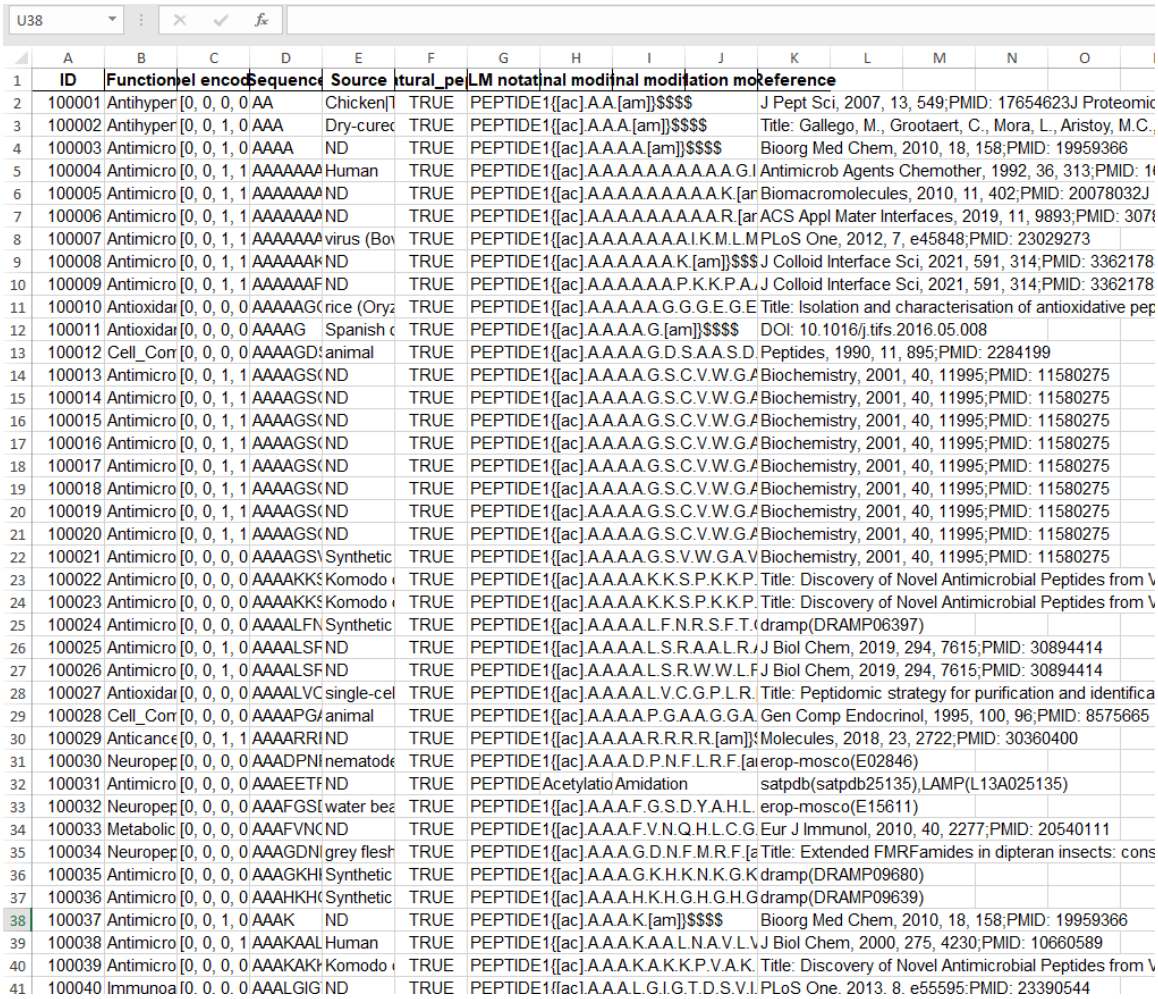
Wang, X., Zhang, Y., & Wang, Y. (2021). DeepHL: Deep learning framework for predicting hemolytic activity of peptides. BMC Genomics, 22, 111. https://doi.org/10.1186/s12864-021-07422-w

Wei, L., Zhou, C., Chen, H., Song, J., Su, R., & Zou, Q. (2018). ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anticancer peptides. Bioinformatics, 34(23), 4007–4016. https://doi.org/10.1093/bioinformatics/bty480

Xie, Y., Zhang, M., Li, J., Zhou, T., & Zhao, Y. (2025). BertADP: A transformer-based model for predicting anti-diabetic peptides. *BMC Biology, 23

Appendix-A

Dataset of BioDADpred



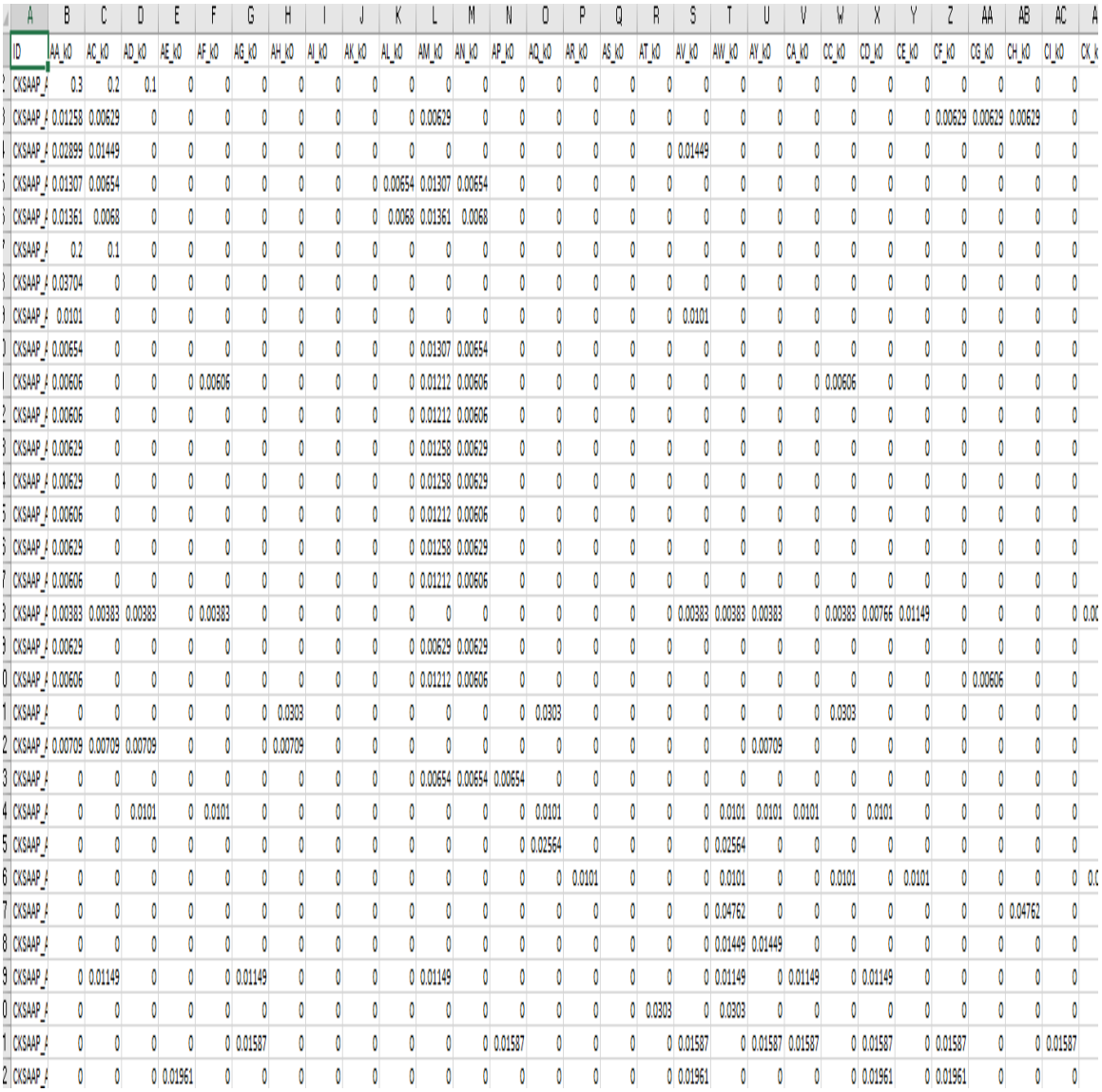
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Function	el encod	Sequence	Source	tural_pe	LM notat	inal modi	inal modi	ation mo	eference				
2	100001	Antihyper	[0, 0, 0, 0	AA	Chicken\T	TRUE	PEPTIDE1{[ac].A.A.[am]}$$$$				J Pept Sci, 2007, 13, 549;PMID: 17654623J Proteomic				
3	100002	Antihyper	[0, 0, 1, 0	AAA	Dry-cured	TRUE	PEPTIDE1{[ac].A.A.A.[am]}$$$$				Title: Gallego, M., Grootaert, C., Mora, L., Aristoy, M.C.,				
4	100003	Antimicro	[0, 0, 1, 0	AAAA	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.[am]}$$$$				Bioorg Med Chem, 2010, 18, 158;PMID: 19959366				
5	100004	Antimicro	[0, 0, 1, 1	AAAAAAA	Human	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.A.A.A.A.G.I				Antimicrob Agents Chemother, 1992, 36, 313;PMID: 1				
6	100005	Antimicro	[0, 0, 1, 1	AAAAAAA	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.A.A.A.K.[ar				Biomacromolecules, 2010, 11, 402;PMID: 20078032J				
7	100006	Antimicro	[0, 0, 1, 1	AAAAAAA	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.A.A.A.R.[ar				ACS Appl Mater Interfaces, 2019, 11, 9893;PMID: 307				
8	100007	Antimicro	[0, 0, 1, 1	AAAAAAA	virus (Bov	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.I.K.M.L.M				PLoS One, 2012, 7, e45848;PMID: 23029273				
9	100008	Antimicro	[0, 0, 1, 1	AAAAAAK	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.K.[am]}$$$				J Colloid Interface Sci, 2021, 591, 314;PMID: 3362178				
10	100009	Antimicro	[0, 0, 1, 1	AAAAAAF	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.A.A.P.K.K.P.A./				J Colloid Interface Sci, 2021, 591, 314;PMID: 3362178				
11	100010	Antioxidar	[0, 0, 0, 0	AAAAAG(	rice (Oryz	TRUE	PEPTIDE1{[ac].A.A.A.A.A.G.G.G.E.G.E				Title: Isolation and characterisation of antioxidative pep				
12	100011	Antioxidar	[0, 0, 0, 0	AAAAG	Spanish c	TRUE	PEPTIDE1{[ac].A.A.A.A.G.[am]}$$$$				DOI: 10.1016/j.tifs.2016.05.008				
13	100012	Cell_Com	[0, 0, 0, 0	AAAAGD:	animal	TRUE	PEPTIDE1{[ac].A.A.A.A.G.D.S.A.A.S.D				Peptides, 1990, 11, 895;PMID: 2284199				
14	100013	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
15	100014	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
16	100015	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
17	100016	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
18	100017	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
19	100018	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
20	100019	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
21	100020	Antimicro	[0, 0, 1, 1	AAAAGS(	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.C.V.W.G.A				Biochemistry, 2001, 40, 11995;PMID: 11580275				
22	100021	Antimicro	[0, 0, 0, 0	AAAAGS\	Synthetic	TRUE	PEPTIDE1{[ac].A.A.A.A.G.S.V.W.G.A.V				Biochemistry, 2001, 40, 11995;PMID: 11580275				
23	100022	Antimicro	[0, 0, 0, 0	AAAAKK:	Komodo (	TRUE	PEPTIDE1{[ac].A.A.A.A.K.K.S.P.K.K.P.				Title: Discovery of Novel Antimicrobial Peptides from V				
24	100023	Antimicro	[0, 0, 0, 0	AAAAKK:	Komodo (	TRUE	PEPTIDE1{[ac].A.A.A.A.K.K.S.P.K.K.P.				Title: Discovery of Novel Antimicrobial Peptides from V				
25	100024	Antimicro	[0, 0, 0, 0	AAAALFN	Synthetic	TRUE	PEPTIDE1{[ac].A.A.A.A.L.F.N.R.S.F.T.(				dramp(DRAMP06397)				
26	100025	Antimicro	[0, 0, 1, 0	AAAALSF	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.L.S.R.A.A.L.R./				J Biol Chem, 2019, 294, 7615;PMID: 30894414				
27	100026	Antimicro	[0, 0, 1, 0	AAAALSF	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.L.S.R.W.W.L.F				J Biol Chem, 2019, 294, 7615;PMID: 30894414				
28	100027	Antioxidar	[0, 0, 0, 0	AAAALVC	single-cel	TRUE	PEPTIDE1{[ac].A.A.A.A.L.V.C.G.P.L.R.				Title: Peptidomic strategy for purification and identifica				
29	100028	Cell_Com	[0, 0, 0, 0	AAAAPG/	animal	TRUE	PEPTIDE1{[ac].A.A.A.A.P.G.A.A.G.G.A.				Gen Comp Endocrinol, 1995, 100, 96;PMID: 8575665				
30	100029	Anticance	[0, 0, 1, 1	AAAARRI	ND	TRUE	PEPTIDE1{[ac].A.A.A.A.R.R.R.R.[am]}$				Molecules, 2018, 23, 2722;PMID: 30360400				
31	100030	Neuropep	[0, 0, 0, 0	AAADPNI	nematode	TRUE	PEPTIDE1{[ac].A.A.A.D.P.N.F.L.R.F.[a				erop-mosco(E02846)				
32	100031	Antimicro	[0, 0, 0, 0	AAAEETF	ND	TRUE	PEPTIDE	Acetylatio	Amidation		satpdb(satpdb25135),LAMP(L13A025135)				
33	100032	Neuropep	[0, 0, 0, 0	AAAFGSI	water bea	TRUE	PEPTIDE1{[ac].A.A.A.F.G.S.D.Y.A.H.L.				erop-mosco(E15611)				
34	100033	Metabolic	[0, 0, 0, 0	AAAFVNC	ND	TRUE	PEPTIDE1{[ac].A.A.A.F.V.N.Q.H.L.C.G.				Eur J Immunol, 2010, 40, 2277;PMID: 20540111				
35	100034	Neuropep	[0, 0, 0, 0	AAAGDNI	grey flesh	TRUE	PEPTIDE1{[ac].A.A.A.G.D.N.F.M.R.F.[a				Title: Extended FMRFamides in dipteran insects: cons				
36	100035	Antimicro	[0, 0, 0, 0	AAAGKHI	Synthetic	TRUE	PEPTIDE1{[ac].A.A.A.G.K.H.K.N.K.G.K				dramp(DRAMP09680)				
37	100036	Antimicro	[0, 0, 0, 0	AAAHKH(	Synthetic	TRUE	PEPTIDE1{[ac].A.A.A.H.K.H.G.H.G.H.G				dramp(DRAMP09639)				
38	100037	Antimicro	[0, 0, 1, 0	AAAK	ND	TRUE	PEPTIDE1{[ac].A.A.A.K.[am]}$$$$				Bioorg Med Chem, 2010, 18, 158;PMID: 19959366				
39	100038	Antimicro	[0, 0, 0, 1	AAAKAAL	Human	TRUE	PEPTIDE1{[ac].A.A.A.K.A.A.L.N.A.V.L.V				J Biol Chem, 2000, 275, 4230;PMID: 10660589				
40	100039	Antimicro	[0, 0, 0, 0	AAAKAKI	Komodo (	TRUE	PEPTIDE1{[ac].A.A.A.K.A.K.K.P.V.A.K.				Title: Discovery of Novel Antimicrobial Peptides from V				
41	100040	Immunoa	[0, 0, 0, 0	AAALGIG	ND	TRUE	PEPTIDE1{[ac].A.A.A.L.G.I.G.T.D.S.V.I.				PLoS One, 2013, 8, e55595;PMID: 23390544				

After filter

	A	B	C	D	E	F
1	ID	sequence	label			
2	ADP_1	AAAAG	1			
3	ADP_2	AAAFVNQ	1			
4	ADP_3	AAALGIGT	1			
5	ADP_4	AAAQHLC	1			
6	ADP_5	AAAQHLC	1			
7	ADP_6	AAATP	1			
8	ADP_7	AAHTSWP	1			
9	ADP_8	AALGIGTD	1			
10	ADP_9	AANPHLC	1			
11	ADP_10	AANQHLC	1			
12	ADP_11	AANQHLC	1			
13	ADP_12	AANQHLC	1			
14	ADP_13	AANQHLC	1			
15	ADP_14	AANQHLC	1			
16	ADP_15	AANQRLC	1			
17	ADP_16	AANRHLC	1			
18	ADP_17	AAPGWPE	1			
19	ADP_18	AAPPQHL	1			
20	ADP_19	AASQHLC	1			
21	ADP_20	ACDGERP	1			
22	ADP_21	ACERLLYP	1			
23	ADP_22	ADPQHLC	1			
24	ADP_23	AEDEVQRI	1			
25	ADP_24	AEDLQVG	1			
26	ADP_25	AEKDEFEH	1			
27	ADP_26	AEKFGPW	1			
28	ADP_27	AELNQLRA	1			
29	ADP_28	AEPNTCAT	1			
30	ADP_29	AEWLHDV	1			
31	ADP_30	AFIEFKAD	1			
32	ADP_31	AFIKATGK	1			
33	ADP_32	AGGGGLD	1			
34	ADP_33	AGGVMTA	1			
35	ADP_34	AGSLQPLA	1			
36	ADP_35	AGSLQPLA	1			
37	ADP_36	AGSLQPLA	1			
38	ADP_37	AHVDKCL	1			
39	ADP_38	AHVQTVG	1			
40	ADP_39	AINSEMFL	1			
41	ADP_40	AKGTTGFE	1			
42	ADP_41	AKGTTGFE	1			
43	ADP_42	AKMHAFT	1			
44	ADP_43	AKSPLF	1			
45	ADP_44	ALADALGS	1			
46	ADP_45	ALEGSLQK	1			
47	ADP_46	ALGDLFQS	1			
48	ADP_47	ALGGA	1			
49	ADP_48	ALIDVFHC	1			
50	ADP_49	ALIPYCVH	1			
51	ADP_50	ALLALWG	1			
1271	NADP_9	FVWQRNI	0			
1272	NADP_10	VITIELSNII	0			
1273	NADP_11	KQRQNKP	0			
1274	NADP_12	NATFYFKII	0			
1275	NADP_13	RAQLKLVA	0			
1276	NADP_14	IIICRKPIIC	0			
1277	NADP_15	MEMVLEL	0			
1278	NADP_16	GFLSILKKY	0			
1279	NADP_17	DAFSPPEA	0			
1280	NADP_18	RKVRGPP	0			
1281	NADP_19	GPYGGGG	0			
1282	NADP_20	WWLSRRI	0			
1283	NADP_21	WYKPAAG	0			
1284	NADP_22	QPELAPED	0			
1285	NADP_23	RSGRGECI	0			
1286	NADP_24	NLDEIDRS	0			
1287	NADP_25	RNLLVGRY	0			
1288	NADP_26	FLGGLLSG	0			
1289	NADP_27	AYWASRM	0			
1290	NADP_28	RRIRPRPP	0			
1291	NADP_29	RKRRKKK	0			
1292	NADP_30	WSSSEVSC	0			
1293	NADP_31	FFFFF	0			
1294	NADP_32	RVISVVQG	0			
1295	NADP_33	LRSFGCRF	0			
1296	NADP_34	GLWNSIK	0			
1297	NADP_35	GYFFFRPR	0			
1298	NADP_36	LEYIDEINL	0			
1299	NADP_37	RQLRIAGF	0			
1300	NADP_38	GLLRRFW	0			
1301	NADP_39	FRKSKEKI	0			
1302	NADP_40	ILFLSIFLCI	0			
1303	NADP_41	CVHAYRA	0			
1304	NADP_42	RRAAVVLI	0			
1305	NADP_43	AHSMIHF	0			
1306	NADP_44	FYDPLVFP	0			
1307	NADP_45	HTTYAADF	0			
1308	NADP_46	KRWWWY	0			
1309	NADP_47	GGPVVMT	0			
1310	NADP_48	GFCWNVC	0			
1311	NADP_49	WWRVVY	0			
1312	NADP_50	LLMQST	0			
1313	NADP_51	RWKKWW	0			
1314	NADP_52	ALYNSEDL	0			
1315	NADP_53	FRRWWK	0			
1316	NADP_54	MVILVFSL	0			
1317	NADP_55	GLMSVLG	0			
1318	NADP_56	GKLWLKG	0			
1319	NADP_57	QRSVSRAA	0			
1320	NADP_58	ATDIPCLL	0			
1321	NADP_59	QPPIRNPF	0			

Appendix-B

After balancing our main Dataset.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	A
	ID	AA_k0	AC_k0	AD_k0	AE_k0	AF_k0	AG_k0	AH_k0	AI_k0	AK_k0	AL_k0	AM_k0	AN_k0	AP_k0	AQ_k0	AR_k0	AS_k0	AT_k0	AV_k0	AW_k0	AY_k0	CA_k0	CC_k0	CD_k0	CE_k0	CF_k0	CG_k0	CH_k0	CI_k0	CK_k
2	CKSAAP_	0.3	0.2	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	CKSAAP_	0.01258	0.00629	0	0	0	0	0	0	0	0	0.00629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00629	0.00629	0.00629	0	
4	CKSAAP_	0.02899	0.01449	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01449	0	0	0	0	0	0	0	0	0	0	0	
5	CKSAAP_	0.01307	0.00654	0	0	0	0	0	0	0	0.00654	0.01307	0.00654	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	CKSAAP_	0.01361	0.0068	0	0	0	0	0	0	0	0.0068	0.01361	0.0068	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	CKSAAP_	0.2	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	CKSAAP_	0.03704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	CKSAAP_	0.0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0101	0	0	0	0	0	0	0	0	0	0	0	
10	CKSAAP_	0.00654	0	0	0	0	0	0	0	0	0	0.01307	0.00654	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	CKSAAP_	0.00606	0	0	0	0.00606	0	0	0	0	0	0.01212	0.00606	0	0	0	0	0	0	0	0	0	0.00606	0	0	0	0	0	0	0	
12	CKSAAP_	0.00606	0	0	0	0	0	0	0	0	0	0.01212	0.00606	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	CKSAAP_	0.00629	0	0	0	0	0	0	0	0	0	0.01258	0.00629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	CKSAAP_	0.00629	0	0	0	0	0	0	0	0	0	0.01258	0.00629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	CKSAAP_	0.00606	0	0	0	0	0	0	0	0	0	0.01212	0.00606	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	CKSAAP_	0.00629	0	0	0	0	0	0	0	0	0	0.01258	0.00629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	CKSAAP_	0.00606	0	0	0	0	0	0	0	0	0	0.01212	0.00606	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	CKSAAP_	0.00383	0.00383	0.00383	0	0.00383	0	0	0	0	0	0	0	0	0	0	0	0	0.00383	0.00383	0.00383	0	0.00383	0.00766	0.01149	0	0	0	0	0.0
19	CKSAAP_	0.00629	0	0	0	0	0	0	0	0	0	0.00629	0.00629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	CKSAAP_	0.00606	0	0	0	0	0	0	0	0	0	0.01212	0.00606	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00606	0	0	
21	CKSAAP_	0	0	0	0	0	0	0.0303	0	0	0	0	0	0	0.0303	0	0	0	0	0	0	0	0.0303	0	0	0	0	0	0	0	
22	CKSAAP_	0.00709	0.00709	0.00709	0	0	0	0.00709	0	0	0	0	0	0	0	0	0	0	0	0	0.00709	0	0	0	0	0	0	0	0	0	
23	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0.00654	0.00654	0.00654	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	CKSAAP_	0	0	0.0101	0	0.0101	0	0	0	0	0	0	0	0	0.0101	0	0	0	0	0.0101	0.0101	0.0101	0	0.0101	0	0	0	0	0	0	
25	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02564	0	0	0	0	0.02564	0	0	0	0	0	0	0	0	0	0	
26	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0101	0	0	0	0.0101	0	0	0.0101	0	0.0101	0	0	0	0	0	0.0
27	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04762	0	0	0	0	0	0	0	0.04762	0		
28	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01449	0.01449	0	0	0	0	0	0	0	0		
29	CKSAAP_	0	0.01149	0	0	0	0.01149	0	0	0	0	0.01149	0	0	0	0	0	0	0	0.01149	0	0.01149	0	0.01149	0	0	0	0	0	
30	CKSAAP_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0303	0	0.0303	0	0	0	0	0	0	0	0	0		
31	CKSAAP_	0	0	0	0	0	0.01587	0	0	0	0	0	0	0.01587	0	0	0	0	0.01587	0	0.01587	0.01587	0	0.01587	0	0.01587	0	0	0.01587		
32	CKSAAP_	0	0	0	0.01961	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01961	0	0	0	0	0.01961	0	0.01961	0	0	0		

Appendix-C

Pseudo code 1: CV test

```
import os, sys, json, gc, math, random, warnings, subprocess
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd

# ----------------------------
# Paths (prefer your uploaded /mnt/data files)
# ----------------------------
if os.path.exists("/mnt/data") and os.path.exists("/mnt/data/AAC_train.csv"):
    BASE = "/mnt/data"
else:
    BASE = "/kaggle/input/biodsdpep-train-test"

DATASETS = ["CKSAAP","AAC", "PseAAC", "DPC", "Merged"]
TRAIN_FILES = {n: f"{BASE}/{n}_train.csv" for n in DATASETS}
TEST_FILES  = {n: f"{BASE}/{n}_test.csv"  for n in DATASETS}

LABEL_COL = "label"
ID_COL    = "ID"   # only in Merged (drop if present)

# ----------------------------
# Repro & TF setup
# ----------------------------
SEED = 42
np.random.seed(SEED); random.seed(SEED); os.environ["PYTHONHASHSEED"] = str(SEED)

import tensorflow as tf
tf.random.set_seed(SEED)

# GPU growth + mixed precision (OOM-safe)
from tensorflow.keras import backend as K
try:
    gpus = tf.config.list_physical_devices('GPU')
    if gpus:
        for g in gpus:
            tf.config.experimental.set_memory_growth(g, True)
    from tensorflow.keras import mixed_precision
    mixed_precision.set_global_policy("mixed_float16")  # memory savings
except Exception:
    pass

def clear_mem():
    K.clear_session(); gc.collect()

# ----------------------------
# Libs
# ----------------------------
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    matthews_corrcoef, cohen_kappa_score, confusion_matrix
)
from tensorflow.keras import layers, models, callbacks, regularizers
from tensorflow.keras.losses import BinaryCrossentropy

# Robust AdamW import
try:
    from tensorflow.keras.optimizers import AdamW
except Exception:
    subprocess.check_call([sys.executable, "-m", "pip", "install", "-q", "tensorflow-addons"])
    from tensorflow_addons.optimizers import AdamW

# ----------------------------
# Training config (same for all datasets/models)
# ----------------------------
EPOCHS       = 160      # raise to 220/300 later if you want more accuracy
K_FOLDS      = 5
TTA_N        = 5
NOISE_STD    = 0.01
BATCH_START  = 128      # auto-shrinks on OOM
BATCH_MIN    = 16
PATIENCE_ES  = 16
PATIENCE_LR  = 6
MIN_LR       = 1e-6
PRED_BS      = 1024
MIXUP_ALPHA  = 0.25
LABEL_SMOOTH = 0.03
LR_INIT      = 1e-3
WD_INIT      = 1e-4
```

```
# ------------------------------
# Data utils
# ------------------------------
def drop_id(df): return df.drop(columns=[ID_COL]) if ID_COL in df.columns else df

def load_xy(train_path, test_path):
    tr = pd.read_csv(train_path)
    te = pd.read_csv(test_path)
    tr = drop_id(tr); te = drop_id(te)
    lcol_tr = LABEL_COL if LABEL_COL in tr.columns else tr.columns[-1]
    lcol_te = LABEL_COL if LABEL_COL in te.columns else te.columns[-1]
    X_train = tr.drop(columns=[lcol_tr]).values.astype(np.float32)
    y_train = tr[lcol_tr].astype(int).values.astype(np.int32)
    X_test  = te.drop(columns=[lcol_te]).values.astype(np.float32)
    y_test  = te[lcol_te].astype(int).values.astype(np.int32)
    return X_train, y_train, X_test, y_test

def standardize_fit_transform(X_tr, X_val, X_te):
    scaler = StandardScaler()
    X_trs  = scaler.fit_transform(X_tr)
    X_vals = scaler.transform(X_val)
    X_tes  = scaler.transform(X_te)
    return X_trs, X_vals, X_tes, scaler

# ------------------------------
# Metrics (ACCURACY-first)
# ------------------------------
def specificity_score_safe(y_true, y_pred_bin):
    cm = confusion_matrix(y_true, y_pred_bin, labels=[0,1])
    tn, fp, fn, tp = cm.ravel()
    return tn / (tn + fp + 1e-15)

def optimize_accuracy_threshold(y_true_val, y_prob_val):
    best_t, best_acc = 0.5, -1
    for t in np.linspace(0, 1, 1001):  # 0.001 step
        yb = (y_prob_val >= t).astype(int)
        acc = accuracy_score(y_true_val, yb)
        if acc > best_acc:
            best_acc, best_t = acc, t
    return float(best_t), float(best_acc)

def evaluate_all(y_true, y_prob, thr):
    y_bin = (y_prob >= thr).astype(int)
    return dict(
        accuracy    = float(accuracy_score(y_true, y_bin)),
        precision   = float(precision_score(y_true, y_bin, zero_division=0)),
        sensitivity = float(recall_score(y_true, y_bin)),
        specificity = float(specificity_score_safe(y_true, y_bin)),
        f1          = float(f1_score(y_true, y_bin, zero_division=0)),
        mcc         = float(matthews_corrcoef(y_true, y_bin)),
        kappa       = float(cohen_kappa_score(y_true, y_bin)),
    )

# ------------------------------
# MixUp (tabular)
# ------------------------------
def mixup(X, y, alpha=MIXUP_ALPHA):
    if alpha <= 0: return X, y
    lam = np.random.beta(alpha, alpha)
    perm = np.random.permutation(len(X))
    return lam*X + (1-lam)*X[perm], lam*y + (1-lam)*y[perm]

def make_batches(X, y, batch, alpha=MIXUP_ALPHA):
    n = len(X); idx = np.arange(n)
    while True:
        np.random.shuffle(idx)
        for i in range(0, n, batch):
            j = idx[i:i+batch]
            xb, yb = X[j], y[j]
            xb, yb = mixup(xb, yb, alpha=alpha)
            yield xb, yb

# ------------------------------
# SWA helpers
# ------------------------------
def average_weights(w_list):
    return [np.mean([w[i] for w in w_list], axis=0) for i in range(len(w_list[0]))]
```

Appendix-D

Pseudo code 2: Independent test

```
import os, sys, gc, math, random, subprocess, warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd

# ------------------------------
# Paths (edit if needed)
# ------------------------------
BASE = "/kaggle/input/biodadpep-train-test"
DATASETS = ["AAC", "DPC", "CKSAAP", "PseAAC", "Merged"]  # edit to subset
TRAIN = {d: f"{BASE}/{d}_train.csv" for d in DATASETS}
TEST  = {d: f"{BASE}/{d}_test.csv"  for d in DATASETS}
LABEL_COL = "label"
ID_COL    = "ID"   # safe to drop if present

# ------------------------------
# Repro & TF setup
# ------------------------------
SEED = 42
np.random.seed(SEED); random.seed(SEED); os.environ["PYTHONHASHSEED"] = str(SEED)

import tensorflow as tf
tf.random.set_seed(SEED)

# GPU growth + mixed precision (OOM-safer)
from tensorflow.keras import backend as K
try:
    for g in tf.config.list_physical_devices('GPU'):
        tf.config.experimental.set_memory_growth(g, True)
    from tensorflow.keras import mixed_precision
    mixed_precision.set_global_policy("mixed_float16")
except Exception:
    pass

def clear_mem():
    K.clear_session(); gc.collect()

# ------------------------------
# Sklearn
# ------------------------------
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    matthews_corrcoef, cohen_kappa_score, confusion_matrix
)

# ------------------------------
# Keras
# ------------------------------
from tensorflow.keras import layers, models, callbacks, regularizers
from tensorflow.keras.losses import BinaryCrossentropy

# Robust AdamW import
try:
    from tensorflow.keras.optimizers import AdamW
except Exception:
    subprocess.check_call([sys.executable, "-m", "pip", "install", "-q", "tensorflow-addons"])
    from tensorflow_addons.optimizers import AdamW

# ------------------------------
# Train config
# ------------------------------
EPOCHS       = 100
VAL_SIZE     = 0.15      # validation split from TRAIN for ES + threshold
BATCH_START  = 128
BATCH_MIN    = 16
PATIENCE_ES  = 16
PATIENCE_LR  = 6
MIN_LR       = 1e-6
LR_INIT      = 1e-3
WD_INIT      = 1e-4
LABEL_SMOOTH = 0.03
MIXUP_ALPHA  = 0.25
PRED_BS      = 1024   # prediction batch size
# NO TTA: single forward pass only
```

201-51-008

ORIGINALITY REPORT

18%	15%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.biorxiv.org Internet Source	1%
3	ouci.dntb.gov.ua Internet Source	1%
4	www.frontiersin.org Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	Güvenilir, Heval Ataş. "Integration and Analysis of Biological Data for Computational Drug Discovery", Middle East Technical University (Turkey), 2024 Publication	<1%
7	peerj.com Internet Source	<1%
8	github.com Internet Source	<1%
9	webs.iiitd.edu.in Internet Source	<1%
10	Submitted to Sello Editorial Student Paper	<1%
11	digibug.ugr.es Internet Source	<1%
12	www.coursehero.com Internet Source	<1%