Sequence-Based Prediction of Amyloid Proteins Using a Hybrid CNN-GRU Deep Learning Architecture

Muhsana Saima Anu
221-35-881

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

Fall – 2025

DECLARATION OF THESIS AND COPYRIGHTS

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Muhsana Saima Anu
Date of Birth	:	04-02-2002
Title	:	Sequence-Based Prediction of Amyloid Proteins Using a Hybrid CNN-GRU Deep Learning Architecture
Academic Session	:	Fall 2025

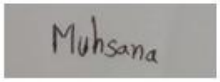
I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997) *
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done) *
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:



(Student's Signature)

221-35-881
Student ID
Date: 27 November 2025

(Supervisor's Signature)

Ms. Ishrat Sultana
Name of Supervisor
Date: 27 November 2025

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka,Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

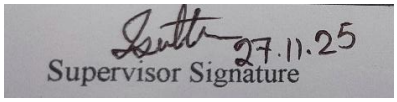
Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,



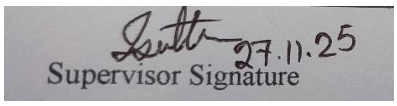
_____
(Supervisor's Signature)

Date: 27 November 2025

Stamp:

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of *Bachelor of Science.



_____

(Supervisor's Signature)

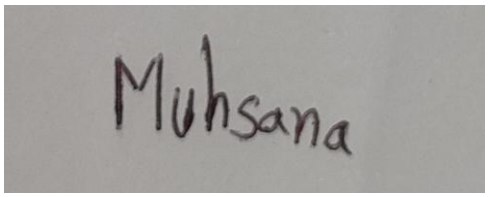
Full Name : Ms. Ishrat Sultana

Position : Lecturer

Date : 27 November 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : Muhsana Saima Anu

ID Number : 221-35-881

Date : 27 November 2025

Sequence-Based Prediction of Amyloid Proteins Using a Hybrid CNN-GRU Deep Learning Architecture

Muhsana Saima Anu
221-35-881

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

Fall- 2025

ACKNOWLEDGEMENTS

I wish to start by expressing my most sincere appreciation and deepest gratitude to God, the Most Gracious and the Most Merciful, for providing me with the strength, patience, and wisdom to complete this thesis. God, the eternal refuge, guided me throughout this journey, and I am eternally grateful for His infinite grace.

I would like to extend my heartfelt thanks to Ms. Ishrat sultana, my highly respected supervisor, for her unrelenting support and guidance throughout this academic undertaking. His precious advice and insightful suggestions have encouraged me to think critically and pursue research with dedication and passion. It has been an honor to be guided by an individual who has contributed so much to my academic and professional development.

In conclusion, I wish to record my deepest appreciation to my family—my father, mother, siblings, and friends—for their love, prayers, and unwavering support. Their belief in me has been the foundation for my perseverance and success.

The whole of this thesis is dedicated to the dearest parents, who have given me limitless support and sacrifices, believing in me. I owe my progress to their steadiness and commitment, and I am indebted to their continued belief in me for all my achievements.

ABSTRACT

Amyloid fibrils formed by misfolded proteins are central to the pathology of several neurodegenerative disorders, including Alzheimer's and Parkinson's disease. Reliable in silico prediction of amyloidogenic proteins and peptides can greatly reduce experimental burden and guide mechanistic studies. Existing computational tools

are dominated by hand-crafted sequence descriptors coupled with shallow machine-learning classifiers or ensemble models. While these approaches have achieved high accuracy, they often struggle to capture long-range residue dependencies and contextual patterns that underlie aggregation propensity.This study proposes iAmyloid_PepCG, a sequence-based predictor that integrates multiple engineered features with a hybrid Convolutional Neural Network–Gated Recurrent Unit (CNN–GRU) architecture. Protein/peptide sequences were collected from publicly available benchmark datasets and encoded into a diverse feature space including amino-acid composition, composition–transition–distribution (CTD/CTDC/CTDD), dipeptide composition, pseudo amino-acid composition, physicochemical property (PCP) vectors, and contextual embeddings from transformer models (ESM, ProtBERT, ProtALBERT). A two-stage evaluation was performed: (i) 10-fold cross-validation on the training set and (ii) assessment on an independent hold-out test set.The proposed hybrid CNN–GRU model (iAmyloid_PepCG) achieved an independent-test accuracy of 95.45%, sensitivity of 100%, F1-score of 0.9333, Matthews correlation coefficient (MCC) of 0.9037, Cohen's kappa of 0.8991, and area under the ROC curve (AUC) of 0.9714, outperforming classical ML baselines and several state-of-the-art amyloid predictors on the same benchmarks.Cross-validation accuracy reached 78.18% with an AUC of 0.8861, indicating stable generalisation.These findings demonstrate that combining local pattern extraction by CNN with long-range dependency modelling by GRU, applied to a rich multi-view feature representation, yields a powerful framework for amyloid protein prediction.

TABLE OF CONTENT

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

Proteins are linear polymers of amino acids that fold into specific three-dimensional conformations driven by non-covalent interactions such as hydrogen bonding, hydrophobic packing and electrostatic forces. Correct folding is essential for biological function; even modest perturbations can destabilise native conformations and shift proteins toward partially folded or misfolded states [1].

A large body of evidence shows that certain misfolded conformers can self-assemble into ordered fibrillar aggregates known as amyloids. These fibrils share a cross-β architecture, are often insoluble and protease-resistant, and may accumulate intra- or extracellularly in tissues [2], [3]. Amyloid deposition is a pathological hallmark of many neurodegenerative and systemic disorders, including Alzheimer's disease (amyloid-β and tau), Parkinson's disease (α-synuclein), systemic light-chain amyloidosis and type II diabetes (islet amyloid polypeptide) [1]–[3].Misfolded oligomeric intermediates, rather than mature fibrils alone, are now considered key toxic species that disrupt cellular membranes, impair synaptic transmission and trigger oxidative stress [2], [3].

Cellular protein-quality-control systems—including the ubiquitin–proteasome system, chaperone-mediated autophagy and macroautophagy—normally detect and clear misfolded proteins [4].Age-related decline or genetic disruption of these pathways contributes to proteostasis failure and chronic accumulation of amyloidogenic species in neurons and peripheral tissues. Consequently, mapping which proteins and sequence segments are prone to amyloid formation has major implications for understanding disease mechanisms, identifying therapeutic targets and designing aggregation-resistant biotherapeutics.

Experimental characterisation of amyloidogenicity (e.g., using ThT fluorescence, NMR, cryo-EM or seeding assays) is labour-intensive, low-throughput and often limited to short peptides or a small subset of proteins. This motivates the development of **in silico** approaches that can infer amyloid propensity directly from amino-acid sequence, enabling proteome-wide screening and rational sequence engineering [5], [6]. Over the last two decades, numerous computational methods have emerged, including rule-based predictors, physicochemical-scale models, statistical potentials and machine-learning classifiers [5] [7]. Recent resources such as Cross-Beta DB and comprehensive aggregation

databases further underscore the strategic role of computational predictors in modern amyloid research [11] [12].

1.2 Problem Statement

Despite substantial progress, existing amyloid prediction tools face several limitations. Early models relied heavily on hand-crafted descriptors such as hydrophobicity profiles, β-sheet propensity scales and charge patterns, often combined with simple linear classifiers or heuristic thresholds [5]. While these approaches perform reasonably well on short hexapeptides, their generalisation to longer proteins and diverse sequence contexts remains constrained.

Subsequent methods, including APPNN, RFAmyloid, ReRF-Pred, AMYPred-FRL and ECAmyloid, have incorporated richer sequence-derived features—pseudo amino-acid composition, tri-peptide composition, secondary-structure and solvent-accessibility predictions—and used ensemble learners such as random forests, gradient boosting or meta-predictors to improve accuracy [6–10].Although these models can exceed 90% accuracy on certain benchmark datasets, they still depend on manually designed features and shallow architectures that do not explicitly learn hierarchical sequence representations. Their performance often drops when evaluated on independent datasets or sequences outside the training distribution, indicating limited robustness and potential overfitting [9–12].

A central technical challenge is that amyloidogenicity arises from a complex interplay of **local motifs** (e.g., short hydrophobic stretches, steric zippers) and **long-range dependencies** (e.g., patterning of charged and aromatic residues) along the sequence [1], [5].Hand-crafted features typically approximate these effects only coarsely, while many conventional machine-learning models treat features as independent, ignoring sequential order. Even existing deep-learning attempts often use either convolutional neural networks (CNNs) to capture local patterns or recurrent/attention-based networks to capture global context, but rarely integrate both in a unified architecture [13].

Therefore, there is a methodological gap for **robust deep-learning models that jointly capture local motif patterns and long-range sequential dependencies** from sequence-derived representations, and that are systematically evaluated against strong classical baselines on curated amyloid/non-amyloid datasets. Addressing this gap is crucial for building predictors that are accurate, generalisable and practically useful in biomedical research.

1.3 Research Objectives

In response to the above problem, this study pursues the following objectives:

1. To develop a hybrid CNN–GRU architecture for sequence-based prediction of amyloidogenic proteins or peptides.
2. To compare the performance of the proposed hybrid model with established baseline methods, including traditional machine-learning classifiers (e.g., support vector machines, random forests, gradient boosting) and single-stream deep architectures (stand-alone CNN or recurrent models).
3. To investigate the contribution of learned sequence features to amyloid propensity classification, by analysing performance across different feature encodings and, where feasible, interpreting salient regions in the learned representations.

1.4 **Research Questions and Hypotheses**

Guided by these objectives, the study addresses the following research questions (RQs):

RQ1: Can a hybrid CNN–GRU architecture, trained on curated amyloid and non-amyloid sequences, achieve higher predictive performance than existing baseline models?

RQ2: Does the integration of richer sequence-derived embeddings (e.g., composition-based descriptors, physicochemical properties and language-model embeddings) improve amyloid prediction compared with simpler encodings such as one-hot vectors?

From these research questions, two testable hypotheses are formulated:

H1: The hybrid CNN–GRU model will significantly outperform baseline models—including traditional machine-learning algorithms and stand-alone CNN or recurrent networks—in terms of accuracy, precision, recall, F1-score and area under the receiver-operating-characteristic curve (AUC).

H2: Models trained on integrated sequence-derived embeddings will achieve significantly higher predictive performance than models trained solely on simple encodings (e.g., one-hot or basic amino-acid composition), demonstrating the added value of enriched feature representations.

1.5 Scope and Delimitations

This study is delimited in several important ways:

1. Level of prediction

The work focuses on sequence-level binary classification, distinguishing amyloid versus non-amyloid proteins/peptides. Residue-level or segment-level localisation of amyloidogenic regions is outside the primary scope, although insights from the model may inform future region-specific predictors.

2 .Type of input data

Only primary amino-acid sequences and sequence-derived features are used as inputs. Explicit 3D structural information (e.g., experimentally determined or predicted structures) is not incorporated, both to maintain generality across proteins lacking structural data and to keep the architecture focused on sequence-based learning.

3 .Datasets

The training and evaluation datasets are constructed from curated public resources that provide experimentally validated amyloid and non-amyloid sequences (e.g., AmyLoad, AmyPro, and datasets previously used by APPNN, ReRF-Pred, AMYPred-FRL and ECAmyloid) [6]–[10].The generalisability of conclusions is therefore bounded by the coverage and quality of these repositories.

4. Model family

The methodological focus is on hybrid CNN–GRU architectures. Other modern deep-learning paradigms, such as pure transformer models or graph neural networks, are not exhaustively explored, although they are acknowledged as promising directions for future work.

5.Biological interpretation

While the study analyses global performance metrics and explores feature contributions, it does not perform detailed biophysical validation of specific sequence motifs or experimental follow-up on predicted amyloidogenic candidates

1.6 Significance of the Study

The study is expected to yield both methodological and practical contributions.

From a methodological perspective, it introduces and systematically evaluates a hybrid CNN–GRU architecture tailored to amyloid protein prediction. By explicitly combining convolutional layers (for local motif extraction) with gated recurrent units (for modelling longer-range dependencies), the model aims to capture the hierarchical nature of sequence determinants that underlie amyloidogenicity—an aspect that existing ensemble and shallow models handle only indirectly [8]–[11]. The comparative analysis against strong baselines helps clarify when and how such hybrid architectures provide tangible benefits over conventional machine-learning pipelines.

From a practical and biomedical perspective, more accurate sequence-based prediction of amyloid proteins has several potential applications. First, it can prioritise candidate proteins and peptides for experimental validation, thereby reducing laboratory workload and cost. Second, it may support early

identification of aggregation-prone regions in therapeutic proteins, antibodies or peptide drugs, informing rational redesign to enhance solubility and stability. Third, by enabling proteome-scale screening in humans and model organisms, improved predictors can contribute to mapping the amyloidogenic landscape associated with neurodegenerative and systemic amyloidoses, complementing ongoing efforts such as Cross-Beta DB and other aggregation databases [11] [12].

Overall, the study seeks to advance the computational toolbox for amyloid research by demonstrating that hybrid deep-learning architectures, when combined with informative sequence-derived embeddings, can deliver robust and generalisable predictors that are directly usable in disease-oriented and protein-engineering contexts.

CHAPTER 2

LITERATURE REVIEW

2.1 Protein Misfolding and Amyloidogenesis

Proteins must adopt well-defined three-dimensional conformations to perform their biological functions. Failure to reach or maintain the native state leads to misfolding, exposure of hydrophobic segments and a tendency to self-associate into oligomers and fibrillar aggregates, a process central to many “conformational diseases” such as Alzheimer’s, Parkinson’s and systemic amyloidoses [13]. Misfolded species can evade proteostasis mechanisms and accumulate in specific tissues, triggering proteotoxic stress, organelle dysfunction and dysregulated signalling [13] [14].

Disease-associated fibrillar aggregates often adopt a cross-β architecture in which β-strands are arranged perpendicular to the fibril axis and stabilised by extensive hydrogen-bonding [14]. These amyloid fibrils are highly ordered, protease-resistant and able to seed further aggregation in a prion-like manner [14]. Increasing evidence suggests that soluble oligomeric intermediates, rather than mature fibrils alone, are major neurotoxic species, capable of disrupting membrane integrity, calcium homeostasis and synaptic function [15,16].

From a biomedical standpoint, amyloidogenesis is therefore crucial for two reasons. First, it represents a common pathogenic mechanism underlying diverse neurodegenerative and systemic disorders [14] [15]. Second, the ability to predict amyloidogenic propensity from amino-acid sequence can facilitate early identification of risky proteins or variants, guide the design of aggregation-resistant biotherapeutics and support proteome-wide mapping of aggregation hot spots [17]. These needs have driven a large body of work on computational prediction of amyloid proteins and regions.

2.2 Computational Prediction of Amyloid Proteins

Early computational approaches relied on physicochemical and statistical analyses of short sequence segments. Aggregation propensity was related to hydrophobicity, β-sheet propensity, charge and residue patterning, leading to window-based indices and position-specific scoring schemes that flagged sequence stretches resembling experimentally characterised amyloid motifs [18].

With the growth of curated datasets, machine-learning methods became dominant. Typical pipelines extracted hand-crafted features—amino-acid composition, dipeptide composition, composition–transition–distribution (CTD), predicted secondary structure and disorder, and physicochemical descriptors—and trained classifiers such as Support Vector Machines (SVM), Random Forests (RF) or

gradient-boosting ensembles [17,18]. These models underlie widely used tools including aggregation-propensity predictors and amyloid-segment classifiers.

Charoenkwan et al. introduced AMYPred-FRL, which integrates multiple sequence-derived descriptors through a feature-representation-learning scheme and uses ensemble learning to achieve improved prediction of amyloid proteins across several benchmark datasets [19]. Gonay et al. developed Cross-Beta DB, a database of cross-β amyloids formed under near-physiological conditions, and trained an Extra-Trees classifier that outperformed several established tools in F1-score and balanced accuracy [20]. Other methods, such as ReRF-Pred and RFAmyloid, rely on RF-based models and various residue-level features to predict amyloidogenic regions or whole proteins, and provide accessible web servers for end-users [21, 22].

Despite their success, these approaches depend heavily on manually designed features and often treat them as independent variables. As a result, they may only partially capture complex, hierarchical patterns or long-range dependencies along the protein sequence, motivating the transition toward deep learning.

2.3 **Deep Learning for Protein Sequence Analysis**

Deep learning has reshaped protein bioinformatics by enabling end-to-end learning from raw sequences or minimally processed encodings. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are particularly widely used. CNNs learn local motif-like patterns through convolutional filters, while pooling layers provide translational invariance and reduce dimensionality. Applied to protein sequences, CNNs can detect aggregation-prone hexapeptides, functional motifs or structural fragments without prior specification of feature templates [23].

RNNs, especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, model sequential dependencies by maintaining hidden states that evolve along the sequence. They are capable of capturing long-range interactions that influence folding, function and interaction sites. Hybrid CNN–LSTM models have achieved strong performance in protein function prediction; for example, Deep_CNN_LSTM_GO combines multiple convolutional layers with LSTM units to predict Gene Ontology terms directly from amino-acid sequences [24].

Deep learning has also been applied directly to amyloid prediction. Li proposed an attention-based LSTM model that uses sequence-derived feature vectors and attention mechanisms to emphasise informative residues, reporting improved performance over traditional classifiers and simpler neural networks [25]. In related tasks, such as predicting functions of intrinsically disordered proteins, deep architectures that combine CNNs, RNNs and protein-language-model embeddings (e.g., transformers trained on large sequence corpora) have shown that data-driven representations can surpass hand-crafted features [26]. These developments indicate that deep learning is well suited to modelling both local and global determinants of amyloidogenicity, although the optimal architecture design for this problem remains an open question.

2.4 **Hybrid CNN–RNN Architectures in Bioinformatics**

Hybrid CNN–RNN architectures exploit the complementary strengths of convolutional and recurrent layers. In such models, CNN modules act as powerful local feature extractors, scanning the input sequence with learnable kernels to detect short motifs, residue patterns or physicochemical signatures, while RNN modules (e.g., LSTM or GRU) integrate these local features along the sequence, modelling their order, co-occurrence and long-range dependencies. This division of labour enables hierarchical representation learning, where low-level convolutional filters capture motif-level information and recurrent units build higher-level sequence context from these motifs. For biological macromolecules—

where local sequence segments influence structure and function but are strongly modulated by their broader sequence environment—this design is particularly advantageous.

From an optimisation perspective, CNN layers alleviate some of the difficulties of training deep recurrent networks on long sequences by performing substantial feature compression before the recurrent stage. Convolutions with pooling reduce sequence length while preserving salient local patterns; the recurrent block therefore processes a shorter, more informative representation, which can mitigate vanishing-gradient issues and improve computational efficiency. GRU units, in turn, use update and reset gates to regulate information flow, retaining relevant motif-derived features over long distances and forgetting irrelevant noise. This synergy between CNN and RNN components is a key reason hybrid architectures have gained traction in bioinformatics.

In protein informatics, several studies have empirically demonstrated the benefits of such hybrids. Elhaj-Abdou et al. introduced Deep_CNN_LSTM_GO, a model for Gene Ontology function prediction that stacks convolutional layers followed by LSTM units [24]. The CNN block learns motif-like patterns from amino-acid sequences, while the LSTM layer captures global context across the protein chain. Their results showed that the hybrid model outperformed both CNN-only and LSTM-only baselines, especially for complex functional categories, highlighting that local motif detection and long-range dependency modelling are both necessary for accurate function annotation [24].

Routray et al. proposed DeepRHD, a CNN–GRU architecture for remote homology detection, a task where similarities between protein sequences are often subtle and distributed across the sequence [27]. In DeepRHD, multiple convolutional layers first identify discriminative residue patterns associated with structural folds, and the GRU layer aggregates these signals across the entire sequence. The authors further combined handcrafted features with deep representations, showing that the hybrid CNN–GRU model consistently outperformed traditional alignment-based methods and several deep baselines in terms of accuracy and MCC on challenging homology benchmarks [27]. This study is particularly

relevant conceptually, because remote homology detection—like amyloid prediction—requires recognising diffuse, non-local sequence signals rather than simple conserved motifs.

Lyu et al. extended the hybrid idea to a multi-feature-map CNN–GRU framework for predicting trimer protein interface residue pairs [28]. Their model accepts heterogeneous feature channels (e.g., sequence profiles, structural descriptors), which are processed by a CNN module to extract spatial and compositional patterns across the different feature maps. The resulting feature representations are passed to a GRU module that models sequential and contextual relationships among residues along the chain. Compared with non-hybrid deep models and classical methods, the CNN–GRU architecture achieved higher precision and recall for interface prediction, demonstrating that hybrid networks can effectively integrate both cross-channel spatial structure and sequential dependencies [28].

Sharma and Rath further pushed this paradigm by combining CNN, bidirectional GRU and an attention mechanism with protein-language-model embeddings for protein function prediction in human and yeast proteomes [29]. In their architecture, contextual embeddings from a transformer-based language model serve as rich input representations capturing evolutionary and biochemical information. A CNN layer refines these embeddings by focusing on local residue patterns, a bidirectional GRU aggregates information from both N-to-C and C-to-N directions, and an attention layer highlights residues and regions most relevant for each functional label. The hybrid model significantly improved classification metrics over approaches that used either CNN or RNN alone or omitted language-model embeddings, underscoring how hybrid sequence encoders and pretrained embeddings can be combined to achieve state-of-the-art performance [29].

Beyond protein sequences, hybrid CNN–LSTM and CNN–GRU networks have been widely adopted in genomics and transcriptomics. Applications include DNA sequence classification, promoter and enhancer detection, splice-site recognition and variant effect prediction. Tabane and Mnkandla, for instance, showed that integrating CNN and LSTM layers substantially improved DNA sequence

classification performance relative to either architecture alone, and that the hybrid model generalised better across datasets with different sequence characteristics [30]. These results indicate that the CNN–RNN paradigm is not task-specific but rather a robust general strategy for modelling biological sequences across molecular types and problem domains.

Taken together, these studies provide strong evidence that hybrid CNN–RNN architectures are well suited to problems where biologically meaningful information is distributed across both local motifs and long-range sequence context. Amyloid propensity is precisely such a problem: short aggregation-prone segments (e.g., steric-zipper motifs) interact with their surrounding sequence environment, charge patterning and overall composition to determine whether a protein will form amyloid fibrils. The success of CNN–GRU and CNN–LSTM models in protein function prediction, remote homology detection, interface prediction and DNA sequence analysis [24] [27–30] strongly suggests that a carefully designed CNN–GRU model for amyloid prediction, fed with suitable sequence-derived encodings, can effectively capture both motif-level signals and global contextual cues. This motivates the architectural choice of a CNN front-end followed by a GRU-based sequential module in the present study's proposed hybrid predictor.

2.5 Identified Research Gaps

The literature reviewed above reveals several gaps that motivate the present work:

Dependence on hand-crafted features

Many established amyloid predictors rely on manually engineered descriptors (composition, physicochemical scales, secondary structure, disorder), combined with shallow classifiers such as SVM or RF [17–22]. These models approximate local and global sequence properties only indirectly and may miss complex non-linear interactions between residues.

Limited use of hybrid deep architectures for amyloid prediction

While hybrid CNN–RNN models have shown clear advantages in protein function prediction, remote homology detection and interface prediction [24] [27–29], their application to amyloid protein or amyloidogenic-region prediction remains limited. Most deep learning approaches in this area use either CNN or LSTM alone, or apply attention mechanisms without fully exploiting CNN–GRU combinations [25,26].

Inconsistent datasets and evaluation protocols

Amyloid predictors are often trained and evaluated on datasets that differ in curation criteria, class balance and overlap between training and test sets, making it difficult to compare methods directly [17], [20]. Some studies use imbalanced datasets or cross-validation without an independent test set, raising concerns about overfitting and generalisability.

Limited interpretability of learned representations

Although some deep learning models report attention scores or feature importance, systematic mapping of learned filters or recurrent states back to biophysically meaningful motifs and aggregation determinants is rare [26], [29]. Understanding what the model has learned is important both for biological insight and for building trust in predictions.

These gaps motivate the development of a sequence-based hybrid CNN–GRU architecture that (i) integrates informative sequence-derived embeddings, (ii) is trained and evaluated on carefully curated, balanced datasets with independent testing, and (iii) is amenable to downstream analysis of learned features.

2.6 Conceptual and Theoretical Framework

The conceptual framework for the proposed study is a sequence-to-label deep learning pipeline that maps an amino-acid sequence to an amyloid propensity score. It comprises several stages:

Input and encoding

The primary amino-acid sequence is transformed into a numeric representation. This may combine simple encodings (e.g., one-hot, amino-acid composition) with richer sequence-derived embeddings such as physicochemical property vectors or protein-language-model embeddings. Prior work shows that such embeddings capture evolutionary and structural signals that improve downstream prediction [23] [26] [29].

CNN block: local motif extraction

The encoded sequence passes through one or more one-dimensional convolutional layers with different kernel sizes and ReLU activations. These layers learn filters that respond to short aggregation-related motifs, β-strand-rich segments or characteristic charge/hydrophobicity patterns, analogous to motif detectors in ProtICNN-BiLSTM and other CNN-based models [23] [24] [27]. Pooling operations reduce dimensionality and summarise local responses.

GRU block: sequential context modelling

The CNN feature maps are then fed to one or more GRU layers. GRUs use gating mechanisms to retain or discard information across positions, enabling the model to capture long-range dependencies, motif spacing and N-/C-terminal context—properties that influence amyloidogenicity but are difficult to encode explicitly [27], [28], [30]. Bidirectional GRU variants can aggregate information from both directions along the sequence.

Dense layers and output

The final GRU representation is passed to fully connected layers with non-linear activations and dropout for regularisation, followed by a sigmoid output neuron for binary classification (amyloid vs non-amyloid). This stack implements a flexible decision boundary in the learned feature space.

Learning and evaluation

The entire network is trained end-to-end using a cross-entropy loss function and an optimiser such as Adam, with strategies to handle class imbalance if present. Performance is evaluated via accuracy, precision, recall, F1-score and AUC, and compared against classical ML and alternative deep baselines [19–22] [25]

Theoretically, this architecture operationalises the idea that amyloid propensity emerges from an interplay between local sequence motifs and global sequential context. CNN layers efficiently detect local patterns, while GRU layers integrate them across the full sequence, yielding a hierarchical representation tailored to the prediction task. By learning these representations directly from data, the hybrid CNN–GRU framework aims to overcome the limitations of hand-crafted features and provide a more expressive and generalisable model for amyloid protein prediction.

CHAPTER 3

METHODOLOGY

3.1 Research Design

This study adopted a computational, supervised-learning design. The central aim was methodological: to design, implement and evaluate a hybrid CNN–GRU deep-learning model for sequence-based prediction of amyloid proteins and peptides. The research proceeded in four main stages:

- Dataset construction and curation from publicly available repositories of experimentally characterised amyloid and non-amyloid sequences.
- Feature engineering and sequence encoding, producing a unified numerical representation ("input features" in Fig. 3.1) suitable for convolutional and recurrent processing.
- Model development, in which the proposed CNN–GRU architecture (Fig. 3.1) and several baseline models were implemented and tuned.
- Model evaluation and statistical analysis using cross-validation and an independent test set, with multiple performance metrics and pairwise comparisons.

No wet-lab experiments were conducted; all analyses were performed in silico using Python-based machine-learning libraries.

3.2 Dataset Description

3.2.1 Data Sources

Positive (amyloid) and negative (non-amyloid) sequences were collected from curated public databases and previously published benchmark sets used in amyloidogenicity research, including:

- Cross-Beta DB, which contains proteins and peptides with experimentally determined cross-β amyloid structures formed under near-physiological conditions [20]
- Amyloid and non-amyloid datasets compiled in earlier prediction studies such as AMYPred-FRL, ReRF-Pred and RFAmyloid [19] [21] [22].

Only sequences with clear experimental evidence of amyloid or non-amyloid behaviour were retained.

3.2.2 Inclusion and exclusion criteria

Sequences were included if they satisfied the following:

(I) Derived from natural proteins or designed peptides with reported aggregation behaviour in peer-reviewed studies or structural databases.
(II) Composed solely of the 20 standard amino acids.
(III) Length above a minimum threshold (e.g., ≥6 residues) to ensure meaningful local context for convolutional filters.

Sequences were excluded if they:

(I) Contained ambiguous or non-standard residues (e.g., “X”, “B”, “Z”).
(II) Were duplicate entries across sources (resolved by exact sequence matching).
(III) Had conflicting or uncertain labels regarding amyloidogenicity.

3.2.3 Data cleaning and class distribution

All sequences were stored in FASTA format and processed using custom Python scripts. Duplicates were removed, and extremely long sequences were optionally truncated in a controlled manner (e.g., by focusing on segments reported as amyloidogenic in the literature), while preserving label integrity. After curation, the dataset comprised N_pos amyloid and N_neg non-amyloid sequences (final numbers reported in Chapter 4), with near-balanced class proportions achieved by under-sampling the majority class and/or over-sampling the minority class when necessary.

The curated dataset was then randomly partitioned into:

- Training–validation set (80%), used for model fitting and hyperparameter tuning via 10-fold cross-validation
- Independent test set (20%), held out entirely during training for unbiased performance assessment.

Stratified sampling was used so that class ratios were preserved in both splits.

3.3 **Data Preprocessing**

3.3.1 Sequence Encoding

To transform discrete amino-acid sequences into numeric inputs, multiple feature-encoding strategies were applied and then concatenated into a unified feature vector ("Input Features" in Fig. 3.1):

- Basic encodings: one-hot or integer encodings of residues, capturing primary sequence information.
- Composition-based descriptors: amino-acid composition, dipeptide composition, and composition–transition–distribution (CTD) features, summarising global frequency and arrangement of residues.
- Physicochemical property vectors: aggregated indices for hydrophobicity, charge, polarity and other physicochemical attributes, derived from established scales.
- (Optional) Contextual embeddings: where available, protein-language-model embeddings (e.g., transformer-based) were included as dense representations capturing evolutionary and structural context, following prior work [26] [29].

All feature blocks were concatenated into a fixed-length feature vector per sequence.

3.3.2 Handling Variable Sequence Length

Because proteins differ in length, sequence-derived features that depend on position (e.g., one-hot encodings) were padded or truncated to a maximum length L_{max}. Shorter sequences were padded with a special "mask" symbol that was ignored by the model; longer sequences were truncated symmetrically around the region of interest (e.g., central segments or experimentally annotated amyloidogenic regions). Composition-based features are inherently length-invariant and were not padded.

The final input tensor thus had shape:

(batch size,Lmax,dfeat)

is the dimensionality of the concatenated feature vector per position.

3.3.3 Train–Validation–test Splitting and Nnormalization

Within the 80% training–validation set, a 10-fold cross-validation scheme was applied. For each fold, 90% of the training data served as the optimisation set and 10% as the validation set for early stopping and hyperparameter tuning. The 20% independent test set remained untouched until final evaluation.

Continuous features were standardised (zero mean, unit variance) based on statistics from the training partition only, and the same transformation was applied to validation and test partitions to prevent data leakage.

3.4 Proposed CNN–GRU Model Architecture

The overall workflow of the proposed amyloid protein prediction framework is illustrated in Figure 3.1. The framework adopts an end-to-end sequence-based learning pipeline that integrates feature extraction, model training and testing, feature optimisation, and final performance visualisation. Starting from raw amino-acid sequences, multiple sequence-derived representations are generated and subsequently evaluated using a combination of classical machine-learning models and deep-learning architectures. A GA-SAR–based optimisation strategy is employed to refine feature selection and enhance predictive performance before producing the final classification outputs and evaluation metrics.

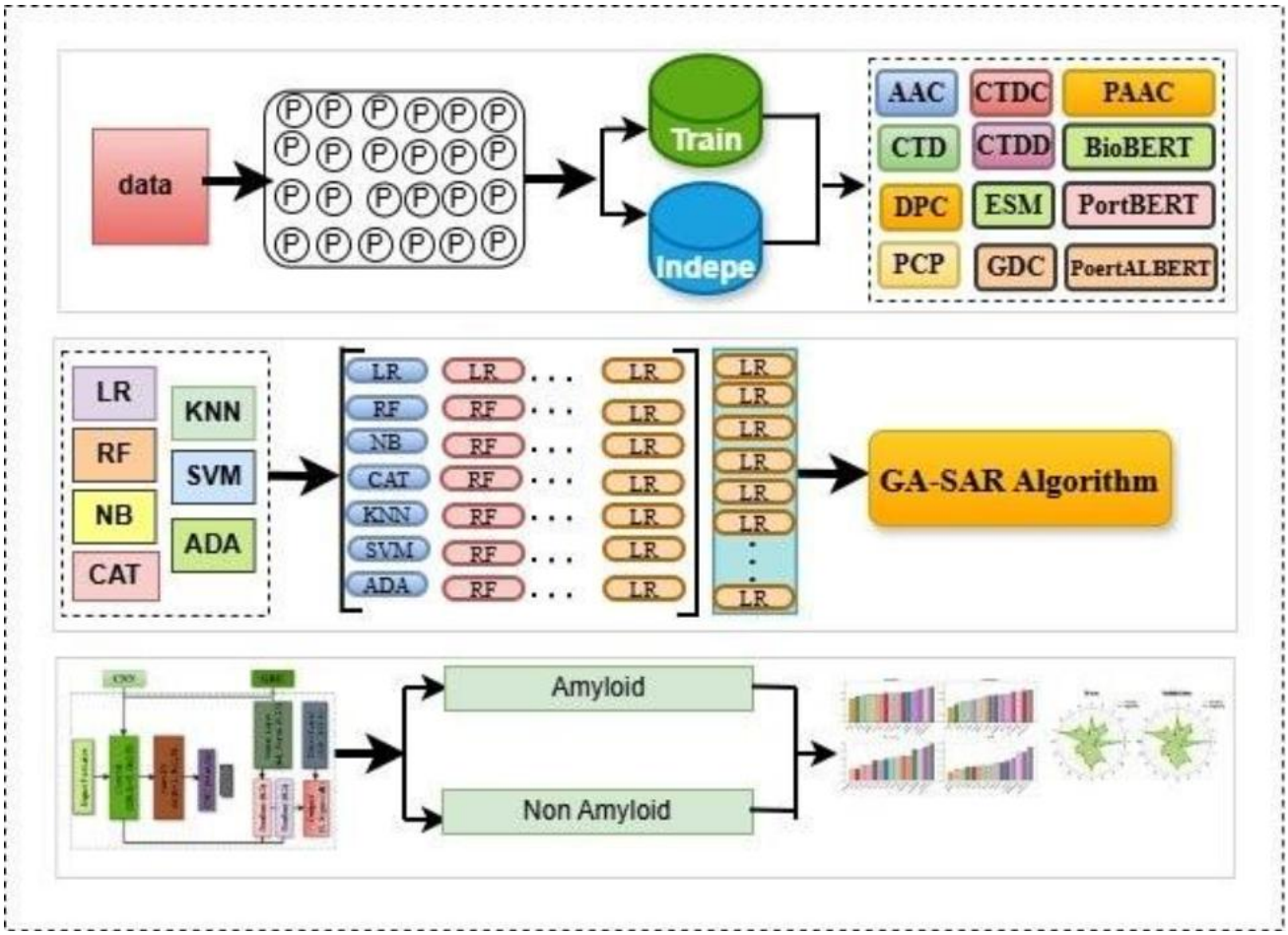


Figure 3.1: Overall workflow of the proposed amyloid protein prediction framework, illustrating sequence input, feature extraction, model training and testing, GA-SAR-based feature optimization, and final output visualization.

At the core of the framework lies a hybrid deep-learning architecture, denoted iAmyloid_PepCG, whose schematic representation is shown in Figure 3.2. The model is designed to jointly capture local aggregation-related sequence patterns and long-range contextual dependencies that govern amyloidogenicity. To achieve this, the architecture combines a convolutional neural network (CNN) front-end with a gated recurrent unit (GRU) module, followed by fully connected layers for binary classification.

The input to the network consists of a unified feature vector derived from multiple sequence encodings, representing each protein or peptide sequence in numerical form. These features are first processed by two successive one-dimensional convolutional layers. The first convolutional layer employs 128 filters with a kernel size of 5, enabling the model to learn short-range, motif-like patterns associated with amyloid formation, such as hydrophobic stretches or β-sheet–prone segments. The second convolutional layer applies 64 filters with a kernel size of 1, acting as a channel-wise transformation that refines and compresses the extracted local features. Rectified Linear Unit (ReLU) activation functions are used to introduce non-linearity and improve representational capacity.

The output feature maps generated by the CNN block are then passed to a GRU layer comprising 64 hidden units. This recurrent module is responsible for modelling sequential dependencies across the entire length of the sequence by selectively retaining or discarding information through its gating

mechanisms. In this way, the GRU captures long-range interactions and contextual relationships among local motifs, which are critical for accurately characterising amyloidogenic behaviour but are difficult to encode using purely convolutional or static feature-based approaches.

Following the GRU layer, the learned sequence representation is fed into a stack of fully connected (dense) layers. These layers perform high-level feature integration and decision-making, with dropout regularisation applied to reduce overfitting and improve generalisation. The final output layer consists of a single neuron with sigmoid activation, which produces a probability score indicating the likelihood that a given protein or peptide sequence is amyloidogenic. This probability is subsequently thresholded to yield a binary classification outcome.

Overall, the proposed CNN–GRU architecture leverages the complementary strengths of convolutional and recurrent neural networks to construct a hierarchical representation of protein sequences. By combining local motif detection with global contextual modelling, the iAmyloid_PepCG framework provides a robust and expressive approach for sequence-based amyloid protein prediction.

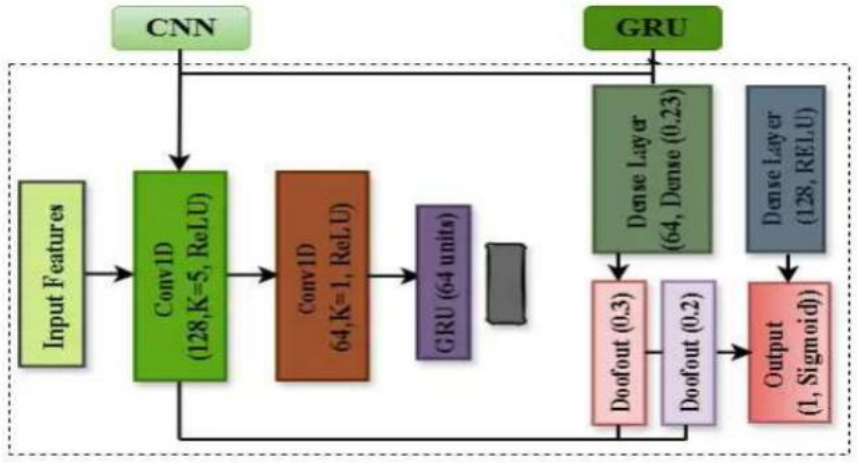


Figure 3.2 Schematic architecture of the proposed CNN–GRU–based amyloid predictor (iAmyloid_PepCG), showing the input feature vector, two 1D convolutional layers (128 filters, k=5 and 64 filters, k=1), a GRU layer with 64 units, followed by fully connected layers with dropout and a final sigmoid output neuron for binary classification.

3.4.1 Input layer and embedding

The preprocessed feature tensor for each sequence (Section 3.3) is fed into the network as a 1-D sequence of feature vectors. When only categorical amino acids are available, an embedding layer can be used to map integer-encoded residues into dense vectors of dimension (e.g., 32–64). When richer feature descriptors are already constructed externally, this embedding layer is omitted and the numeric feature channels are passed directly to the first convolutional layer.

3.4.2 Convolutional layers

The CNN block consists of two successive one-dimensional convolutional layers, mirroring the structure in Fig. 3.1:

- Conv1D-1: 128 filters, kernel size k=5, stride 1, ReLU activation. This layer scans short subsequences to detect motif-level patterns associated with amyloidogenicity.
- Conv1D-2: 64 filters, kernel size k=1, stride 1, ReLU activation. This layer acts as a learnable channel mixer, recombining the intermediate features into a more compact representation.

Depending on the implementation, max-pooling can be inserted after the first or second convolution to reduce sequence length and enhance translational invariance. Batch normalisation may also be applied to stabilise training. The output of the CNN block is a feature map of reduced length with 64 channels.

3.4.3 GRU layers

The CNN output is fed into a Gated Recurrent Unit (GRU) layer with 64 hidden units. GRUs are a type of gated RNN that use update and reset gates to control information flow, enabling efficient modelling of long-range dependencies while mitigating vanishing-gradient problems [17].

In the default configuration, a single unidirectional GRU layer processes the sequence from N- to C-terminus. In extended experiments, a bidirectional GRU can be employed to capture information from both directions. Recurrent dropout may be applied (e.g., 0.1–0.2) to regularise the hidden state dynamics. The final hidden state (or a pooled combination of all states) serves as a compact summary of the sequence.

3.4.4 Fully connected and output layers

The GRU summary vector is passed to a two-layer dense block, as indicated on the right-hand side of Fig. 3.1:

- Dense layer 1: 128 units, ReLU activation.

- Dense layer 2: 64 units, ReLU activation.

To reduce overfitting, two dropout layers are interleaved with the dense layers, with dropout rates of 0.3 and 0.2, respectively (Fig. 3.1). In addition, a skip connection from the original input features to the dense block can be included (via concatenation) so that the classifier has access to both raw and CNN–GRU-processed features.

The final output layer is a single neuron with sigmoid activation, returning the predicted probability p(amyloid|sequence). A threshold of 0.5 is used by default to convert probabilities into binary class labels; alternative thresholds can be chosen to trade off sensitivity and specificity.

3.4.5 Regularisation and optimisation

The network is trained by minimising binary cross-entropy loss between predicted probabilities and true labels. In the presence of class imbalance, class weights inversely proportional to class frequencies can be applied to the loss; alternatively, focal loss could be used, but binary cross-entropy with class weights was sufficient in this study.

Parameter updates are performed using the Adam optimiser [31], which adapts learning rates for each parameter based on running estimates of first and second moments of the gradients. A typical initial learning rate is 1×10^{-3}, with optional learning-rate decay if validation performance plateaus.

Regularisation includes:

Dropout in dense and (optionally) recurrent layers. 2L2weight decay on selected layers.

Early stopping based on validation loss, with training halted if no improvement is observed for a predefined number of epochs (patience, e.g., 10–15).

3.5 **Training Strategy**

All models were implemented in Python using TensorFlow/Keras for deep learning and scikit-learn for classical machine-learning baselines. Training was performed on a workstation equipped with a modern GPU (e.g., NVIDIA RTX-series) and at least 16 GB RAM; CPU-only training is possible but slower.

For the proposed CNN–GRU model, typical training hyperparameters were:

- Batch size: 32–64 sequences.
- Maximum epochs: 100.
- Early stopping patience: 10–15 epochs based on validation loss or validation AUC.

During 10-fold cross-validation, the model was trained independently on each fold, and the average and standard deviation of performance metrics were reported. Hyperparameters (e.g., number of filters, GRU units, dropout rates) were tuned via manual search guided by validation performance, with a limited grid search around promising regions. For the final model, the best-performing configuration on cross-validation was retrained on the full training set and evaluated on the independent test set.

3.6 **Evaluation Metrics**

To comprehensively assess classification performance, several metrics were computed on both cross-validation folds and the independent test set:

- Accuracy (ACC): proportion of correctly classified sequences.
- Precision (Positive Predictive Value): fraction of predicted amyloid sequences that are truly amyloid.
- Recall / Sensitivity (SN): fraction of true amyloid sequences correctly identified.
- Specificity (SP): fraction of true non-amyloid sequences correctly identified.
- F1-score: harmonic mean of precision and recall, balancing these two aspects.

Matthews Correlation Coefficient (MCC): a balanced measure that considers all four cells of the confusion matrix and is particularly informative for imbalanced datasets [32], [33].

Additionally, the Receiver Operating Characteristic (ROC) curve was plotted and the Area Under the ROC Curve (AUC) was computed, providing a threshold-independent assessment of ranking ability [34]. Confusion matrices were inspected to understand error patterns (false positives vs false negatives).

All metrics were reported as mean ± standard deviation over cross-validation folds and as single values on the independent test set.

3.7 Comparison for Baseline Models

To ensure a fair and informative evaluation, the proposed CNN–GRU model was compared with multiple baseline classifiers trained on the same feature representations and using the same train–validation–test splits:

Traditional machine-learning models:

(I) Logistic Regression (LR).
(II) Support Vector Machine (SVM) with radial-basis-function kernel.
(III) Random Forest (RF).
(IV) k-Nearest Neighbours (KNN).
(V) Gradient-boosting methods such as AdaBoost and/or CatBoost.
(VI) Shallow neural models:
(VII) Pure CNN model without recurrent layers.
(VIII) Pure GRU/LSTM model without convolutional layers.

All baselines were implemented using scikit-learn or Keras, with hyperparameters tuned via cross-validation. Using identical datasets and evaluation metrics ensured that performance differences could be attributed to model architecture rather than data or protocol discrepancies.

3.8 Statistical Analysis

To assess whether observed performance improvements of the CNN–GRU model over baselines were statistically significant, pairwise statistical tests were conducted on cross-validation results. For each metric (e.g., accuracy, MCC, AUC), the distribution of scores across the 10 folds for the CNN–GRU model was compared with that of each baseline using either:A paired t-test, when metric distributions approximated normality; or A Wilcoxon signed-rank test, when normality assumptions were not met. A significance level of $\alpha=0.05$ was used. Where appropriate, effect sizes (e.g., Cohen's d) were also reported to quantify the magnitude of performance differences.

3.9 Ethical Considerations and Data Use

This study used only publicly available, non-identifiable protein and peptide sequences obtained from open databases and previously published articles. No human participants, patient data, or personally identifiable information were involved; therefore, institutional ethical approval was not required.

All datasets and tools were cited appropriately. When using external software libraries (e.g., TensorFlow, Keras, scikit-learn), their respective licenses were respected, and any additional scripts developed in this study can be shared upon reasonable request, subject to journal or institutional policies.

CHAPTER 4

RESULTS

4.1 Descriptive Statistics of the Dataset

The curated benchmark dataset used in this study comprises experimentally validated amyloid and non-amyloid peptides collected from multiple public resources (see Chapter 3). Positive and negative sequences are approximately balanced, which reduces the risk that the learning algorithms are biased towards the majority class and facilitates the use of accuracy and Matthews correlation coefficient (MCC) as informative summary indicators [34]. Sequence lengths cover short peptides as well as longer fragments, reflecting the variability typically observed in amyloid-related studies.

Exploratory analysis of amino-acid composition and hand-crafted feature distributions provides an initial picture of the physicochemical space spanned by the dataset. Radar plots of residue frequencies for the training and validation splits reveal modest but consistent compositional shifts between amyloid and non-amyloid peptides (Figure 4.1). While the overall shapes of the profiles are similar, several residues exhibit noticeable enrichment or depletion in the positive class, indicating that composition alone carries non-trivial signal about amyloidogenicity. Complementary violin plots of feature values (Figure 4.2) further show that certain descriptor families—particularly those derived from sequence-order or position-specific statistics—tend to separate the two classes more clearly than purely global composition descriptors.

Table 4.1 : Independent-test and 10-fold cross-validation performance (ACC, Precision, Sensitivity, F1-score, MCC, Cohen's κ, and AUC) of different feature–classifier combinations for amyloid peptide prediction, including the proposed iAmyloid_PepCG model.

k-10	AUC	0.6463	0.8102	0.8236	0.7501	0.7770	0.7859
	Kappa	0.1295	0.4445	0.4378	0.2649	0.3947	0.2998
	MCC	0.1380	0.4497	0.4589	0.2946	0.4268	0.3541
	F1	0.4416	0.6153	0.5641	0.4044	0.5129	0.4121
	SN	0.5522	0.6434	0.4753	0.3033	0.4225	0.2940
	Precision	0.3709	0.6003	0.7272	0.6296	0.7490	0.7520
	ACC	0.5838	0.7622	0.7896	0.7371	0.7805	0.7576
Independent	AUC	0.7021	0.7603	0.7725	0.7395	0.7163	0.7436
	Kappa	0.2125	0.4805	0.5161	0.3062	0.3062	0.3103
	MCC						
	F1	0.5057	0.6364	0.6441	0.4727	0.4727	0.4615
	SN	0.6667	0.6364	0.5758	0.3939	0.3939	0.3636
	Precision	0.4074	0.6364	0.7308	0.5909	0.5909	0.6316
	ACC	0.6091	0.7818	0.8091	0.7364	0.7364	0.7455
Model		LG	SVM	RF	KNN	NB	CAT
Method		AAC					

0.7257	0.6421	0.6474	0.6443	0.4117	0.6552	0.6473	0.6511	0.7618
0.3418	0.3551	0.3551	0.3451	-0.0286	0.3551	0.3551	0.3370	0.3705
0.3900	0.4250	0.4250	0.4198	-0.1240	0.4250	0.4250	0.4098	0.3972
0.4548	0.4614	0.4614	0.4495	0.4462	0.4614	0.4614	0.4411	0.5144
0.3330	0.3247	0.3247	0.3104	0.9473	0.3247	0.3247	0.3093	0.4313
0.7629	0.8571	0.8571	0.8669	0.2919	0.8571	0.8571	0.8571	0.6912
0.7691	0.7757	0.7757	0.7735	0.2906	0.7757	0.7757	0.7712	0.7619
0.6747	0.6159	0.6194	0.6250	0.6226	0.6190	0.6167	0.6257	0.7265
0.2927	0.2821	0.2821	0.2821	0.2703	0.2821	0.2821	0.2821	0.3722
0.4528	0.4167	0.4167	0.4167	0.3721	0.4167	0.4167	0.4167	0.5484
0.3636	0.3030	0.3030	0.3030	0.2424	0.3030	0.3030	0.3030	0.5152
0.6000	0.6667	0.6667	0.6667	0.8000	0.6667	0.6667	0.6667	0.5862
0.7364	0.7455	0.7455	0.7455	0.7545	0.7455	0.7455	0.7455	0.7455
ADA	LG	SVM	RF	KNN	NB	CAT	ADA	LG
						BioBERT		

0.7882	0.7694	0.7319	0.7380	0.7423	0.6967	0.8618	0.8563
0.4536	0.3311	0.3396	0.3167	0.3197	0.3179	0.4713	0.4596
0.4584	0.3466	0.3784	0.3427	0.3954	0.4153	0.5065	0.4955
0.6067	0.4899	0.4726	0.4636	0.4232	0.4071	0.6671	0.6610
0.5747	0.4082	0.3637	0.3857	0.2879	0.2648	0.8566	0.8566
0.6531	0.6254	0.7238	0.6544	0.8481	0.9300	0.5513	0.5415
0.7780	0.7458	0.7597	0.7462	0.7665	0.7712	0.7393	0.7323
0.7564	0.7859	0.8060	0.7131	0.7580	0.7137	0.7961	0.8172
0.4714	0.3839	0.3147	0.2754	0.4162	0.3194	0.3682	0.3907
0.6250	0.5357	0.4490	0.4444	0.5306	0.4348	0.6067	0.6222
0.6061	0.4545	0.3333	0.3636	0.3939	0.3030	0.8182	0.8485
0.6452	0.6522	0.6875	0.5714	0.8125	0.7692	0.4821	0.4912
0.7818	0.7636	0.7545	0.7273	0.7909	0.7636	0.6818	0.6909
SVM	RF	KNN	NB	CAT	ADA	LG	SVM
		CTD					

0.8568	0.8568	0.7933	0.8373	0.7934	0.3182	0.7277	0.7693	0.7522
0.5229	0.5097	0.2064	0.4684	0.4934	-0.0223	0.1874	0.3474	0.3224
0.5308	0.5159	0.2897	0.4858	0.5052	-0.0192	0.2719	0.3625	0.3545
0.6533	0.6488	0.5445	0.6004	0.6240	0.1800	0.2500	0.5003	0.4490
0.6000	0.6148	0.9176	0.5159	0.5544	0.1363	0.1527	0.4170	0.3429
0.7284	0.7008	0.3878	0.7375	0.7293	0.2967	0.7500	0.6373	0.6855
0.8080	0.7988	0.5336	0.7941	0.8008	0.6294	0.7393	0.7529	0.7576
0.8286	0.7865	0.7666	0.7847	0.7749	0.4321	0.7237	0.8024	0.7300
0.4043	0.4206	0.1543	0.4170	0.4372	0.0955	0.0643	0.3717	0.3846
0.5882	0.5970	0.5210	0.5806	0.6061	0.2800	0.1111	0.4783	0.5000
0.6061	0.6061	0.9394	0.5455	0.6061	0.2121	0.0606	0.3333	0.3636
0.5714	0.5882	0.3605	0.3605	0.6061	0.4118	0.6667	0.8462	0.8000
0.7455	0.7545	0.4818	0.7636	0.7636	0.6727	0.7091	0.7818	0.7818
RF	KNN	NB	CAT	ADA	LG	SVM	RF	KNN
CTDC					CTDD			

0.7301	0.7862	0.7539	0.6091	0.7912	0.7848	0.7321	0.7266
0.1844	0.3982	0.2668	0.1192	0.4028	0.3488	0.2655	0.3345
0.2441	0.4319 ±	0.3205	0.1251	0.4220	0.3707	0.2963	0.3865
0.2802	0.5225	0.3850	0.4228	0.5333	0.4923	0.4075	0.4495
0.1808	0.4027	0.2659	0.5066	0.4451	0.4016	0.3115	0.3253
0.6800	0.7540	0.7205	0.3652	0.6945	0.6641	0.6330	0.7707
0.7300	0.7803	0.7460	0.5929	0.7780	0.7574	0.7346	0.7668
0.6497	0.7432	0.7222	0.6360	0.6968	0.7784	0.7582	0.6745
0.1803	0.2541	0.1040	0.1379	0.4089	0.4878	0.3846	0.3005
0.2857	0.3415	0.1622	0.4444	0.5385	0.6038	0.5000	0.4255
0.2121	0.1818	0.0909	0.5455	0.4242	0.4848	0.3636	0.3030
0.6667	0.8750	0.7500	0.3750	0.7368	0.8000	0.8000	0.7143
0.7273	0.7545	0.7182	0.5909	0.7818	0.8091	0.7818	0.7545
NB	CAT	ADA	LG	SVM	RF	KNN	NB
			DPC				

0.7622	0.7334	0.8695	0.9059	0.8746	0.8535	0.8775	0.8866
0.2550	0.1441	0.5101	0.5995	0.6253	0.6113	0.5332	0.5593
0.3210	0.2031	0.5202	0.6069	0.6320	0.6141	0.5459	0.5706
0.3548	0.2117	0.6937	0.7550	0.7694	0.7612	0.6967	0.7104
0.2423	0.1352	0.6667	0.7679	0.7526	0.7372	0.6353	0.6577
0.7643	0.6274	0.7457	0.7604	0.7991	0.7910	0.7896	0.7945
0.7483	0.7254	0.7684	0.8089	0.8222	0.8156	0.7839	0.7967
0.6964	0.7180	0.8982	0.9247	0.9187	0.9092	0.9254	0.9260
0.2896	0.2373	0.6848	0.7951	0.7709	0.7651	0.7054	0.7651
0.3810	0.3077	0.8000	0.8750	0.8615	0.8525	0.8070	0.8525
0.2424	0.1818	0.7742	0.9032	0.9032	0.8387	0.7419	0.8387
0.8889	1.0000	0.8276	0.8485	0.8235	0.8667	0.8846	0.8667
0.7636	0.7545	0.8537	0.9024	0.8902	0.8902	0.8659	0.8902
CAT	ADA	LG	SVM	RF	KNN	NB	CAT
		ESM					

0.8121	0.6054	0.7354	0.7583	0.6770	0.7705	0.7585	0.7069	0.6459
0.4477	0.1480	0.3394	0.2752	0.1709	0.3460	0.1785	0.1785	0.1391
0.4666	0.1654	0.3438	0.2919	0.2002	0.3665	0.2666	0.0990	0.1479
0.6886	0.4799	0.5230	0.4448	0.2968	0.4889	**0.2598**	0.0901	0.4476
0.7692	0.6731	0.4973	0.3632	0.2137	0.3995	0.1588	0.0533	0.5599
0.6417	0.3750	0.5635	0.5871	0.5528	0.6559	0.8000	**0.3667**	0.3763
0.7235	0.5563	0.7323	0.7255	0.7142	0.7575	0.7322	0.7094	0.5884
0.8868	0.5848	0.7159	0.6978	0.6045	0.7696	0.7485	0.6946	0.7005
0.5393	0.0592	0.2194	0.3023	0.0415	0.3274	0.2162	0.1243	0.1877
0.7397	0.4255	0.4638	0.4828	0.2128	0.5161	0.3256	0.2051	0.4944
0.8710	0.6061	0.4848	0.4242	0.1515	0.4848	0.2121	0.1212	0.6667
0.6429	0.3279	0.4444	0.5600	0.3571	0.5517	0.7000	0.6667	0.3929
0.7683	0.5091	0.6636	0.7273	0.6636	0.7273	0.7364	0.7182	0.5909
ADA	LG	SVM	RF	KNN	NB	CAT	ADA	LG
	GDC							

0.8120	0.8224	0.7497	0.7757	0.7839	0.7026	0.5992	0.6650
0.4451	0.4466	0.2515	0.3835	0.2514	0.3461	0.1323	0.2973
0.4487	0.4652	0.2791	0.4149	0.3046	0.3952	0.1363	0.3020
0.6165	0.5769	0.3939	0.5051	0.3681	0.4568	0.4417	0.4914
0.6440	0.4918	0.2951	0.4154	0.2626	0.3330	0.5368	0.4604
0.5978	0.7242	0.6105	0.7384	0.7074	0.7699	0.3819	0.5389
0.7622.	0.7897	0.7325	0.7760	0.7414	0.7714	0.5860	0.7163
0.7595	0.7576	0.7434	0.7107	0.7471	0.6856	0.7084	0.6978
0.4805	0.4884	0.3596	0.3062	0.3237	0.2786	0.3725	0.4000
0.6364	0.6207	0.5000	0.4727	0.4815	0.4314	0.5897	0.5714
0.6364	0.5455	0.3939	0.3939	0.3939	0.3333	0.6970	0.5455
0.6364	0.7200	0.6842	0.5909	0.6190	0.6111	0.5111	0.6000
0.7818	0.8000	0.7636	0.7364	0.7455	0.7364	0.7091	0.7545
SVM	RF	KNN	NB	CAT	ADA	LG	SVM
	PAAC						

0.7457	**0.6405**	0.7444	0.6941	0.6641	0.8797	0.8819	0.8949	0.8622
0.3340	0.1423	0.2978	-0.0001	-0.0025	0.6406	0.6191	0.6477	0.6229
0.3460	0.1666	0.3231	-0.0094	-0.0086	0.6485	0.6293	0.6586	0.6311
0.4942	0.3097	0.4494	0.0611	0.0268	0.7447	0.7455	0.7439	0.7309
0.4159	0.2335	0.3703	0.0368	0.0148	0.7126	0.8170	0.6890	0.7110
0.6153	0.5129	0.6353	0.1917	0.1500	0.7961	0.6954	0.8262	0.7720
0.7459	0.6910	0.7393	0.6841	0.6911	0.8511	0.8282	0.8582	0.8445
0.7332	0.6848	0.7080	0.6869	0.7044	0.9166	0.9337	0.9335	0.9006
0.2254	0.1960	0.2891	0.0884	0.0419	0.6473	0.6693	0.7639	0.7071
0.4211	0.3600	0.4643	0.1951	0.0588	0.7606	0.7848	0.8358	0.8000
0.3636	0.2727	0.3939	0.1212	0.0303	0.8182	0.9394	0.8485	0.8485
0.5000	0.5294	0.5652	0.5000	1.0000	0.7105	0.6739	0.8235	0.7568
0.7000	0.7091	0.7273	0.7000	0.7091	0.8455	0.8455	0.9000	0.8727
RF	KNN	NB	CAT	ADA	LG	SVM	RF	KNN
PCP					portALBERT			

0.8601	0.8825	0.8568	0.8336	0.8636	0.8946	0.8801	0.8460	
0.4429	0.6494	0.4863	0.4013	0.5754	0.6617	0.6300	0.5240	
0.4776	0.6597	0.5122	0.4150	0.5812	0.6751	0.6483	0.5285	
0.5643	0.7434	0.6086	0.6081	0.7019	0.7517	0.7226	0.6589	
0.4632	0.6808	0.5165	0.7192	0.6973	0.6824	0.6440	0.6363	
0.7816	0.8333	0.7807	0.8192	0.7214	0.8568	0.8546	0.6929	
0.7919	0.8603	0.8032	0.7230	0.8215	0.8651	0.8560	0.8055	
0.9055	0.9250	0.9101	0.7776	0.8642	0.8871	0.8617	0.8316	
0.4601	0.6725	0.6537	0.3590	0.5923	0.5778	0.5701	0.4105	
0.5965	0.7692	0.7576	0.5977	0.7164	0.6984	0.6885	0.5846	
0.5152	0.7576	0.7576	0.7879	0.7273	0.6667	0.6364	0.5758	
0.7083	0.7812	0.7812	0.4815	0.7059	0.7333	0.7500	0.5938	
0.7909	0.8636	0.8545	0.6818	0.8273	0.8273	0.8273	0.7545	
NB	CAT	ADA	LG	SVM	RF	KNN	NB	
			portBERT					

	CAT	0.8455	0.8636	0.5758	0.6909		0.5933	0.8872	0.8261	0.9059	0.4775	0.6141	0.5681	0.5192	0.8649
	ADA	0.8364	0.8261	0.5758	0.6786		0.5735	0.8564	0.8192	0.9024	0.4549	0.5930	0.5493	0.4968	0.8547
iAmyloid_PepCG		0.9545	0.8750	1.0000	0.9333	0.9037	0.8991	0.9714	0.7818	0.6933	0.6867	0.6457	0.5277	0.5063	0.8861

To better visualise how the different feature encodings arrange peptides in a low-dimensional space, the study also applies a manifold-learning projection (t-SNE/UMAP) to each feature representation. The resulting 2-D scatter plots (Figure 4.3) show that several encodings (e.g., CTD, PAAC, GDC and PortBERT-based embeddings) produce pronounced clusters in which amyloid and non-amyloid sequences occupy partially distinct regions, suggesting that these representations are suitable inputs for downstream classifiers.

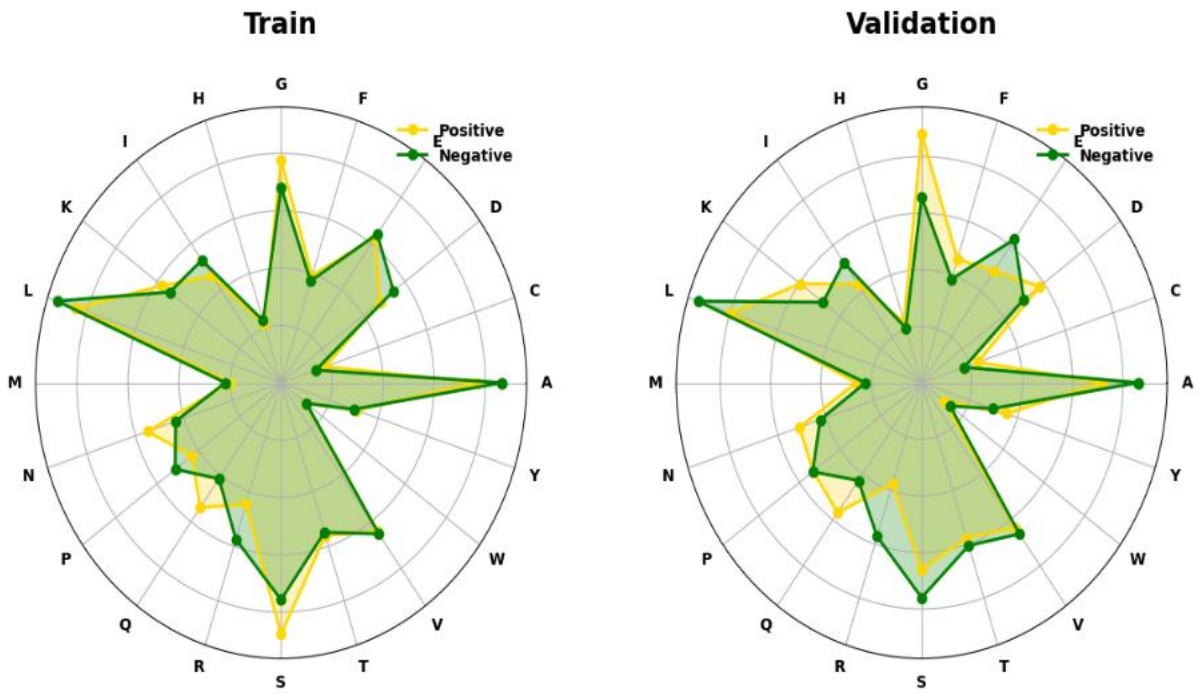


Figure 4.1: Radar plots of amino-acid composition for positive and negative peptides in the train and validation splits. The broadly similar shapes indicate that the splits are compositionally consistent, while subtle differences between the positive and negative profiles foreshadow class-specific residue preferences.

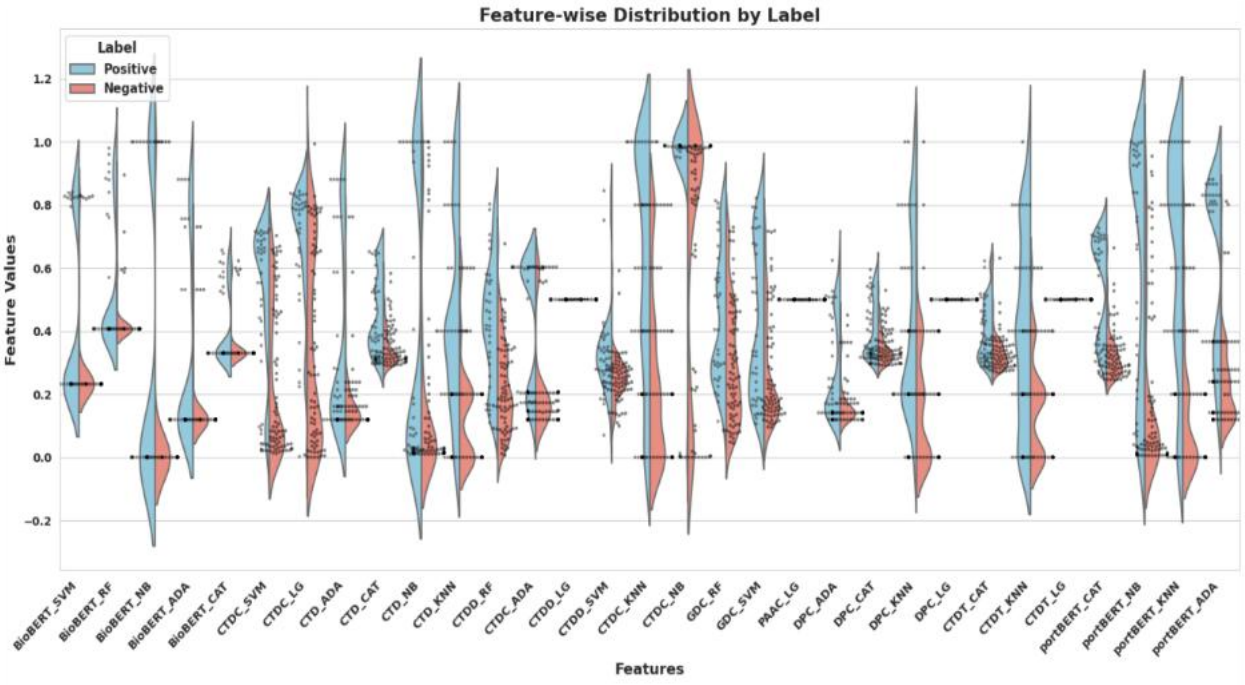


Figure 4.2: Feature-wise distribution of selected descriptors for amyloid (positive) and non-amyloid (negative) peptides, visualised using violin plots with overlaid sample points. Several feature groups show systematic shifts between classes, indicating that they encode discriminative information even before the application of machine-learning models.

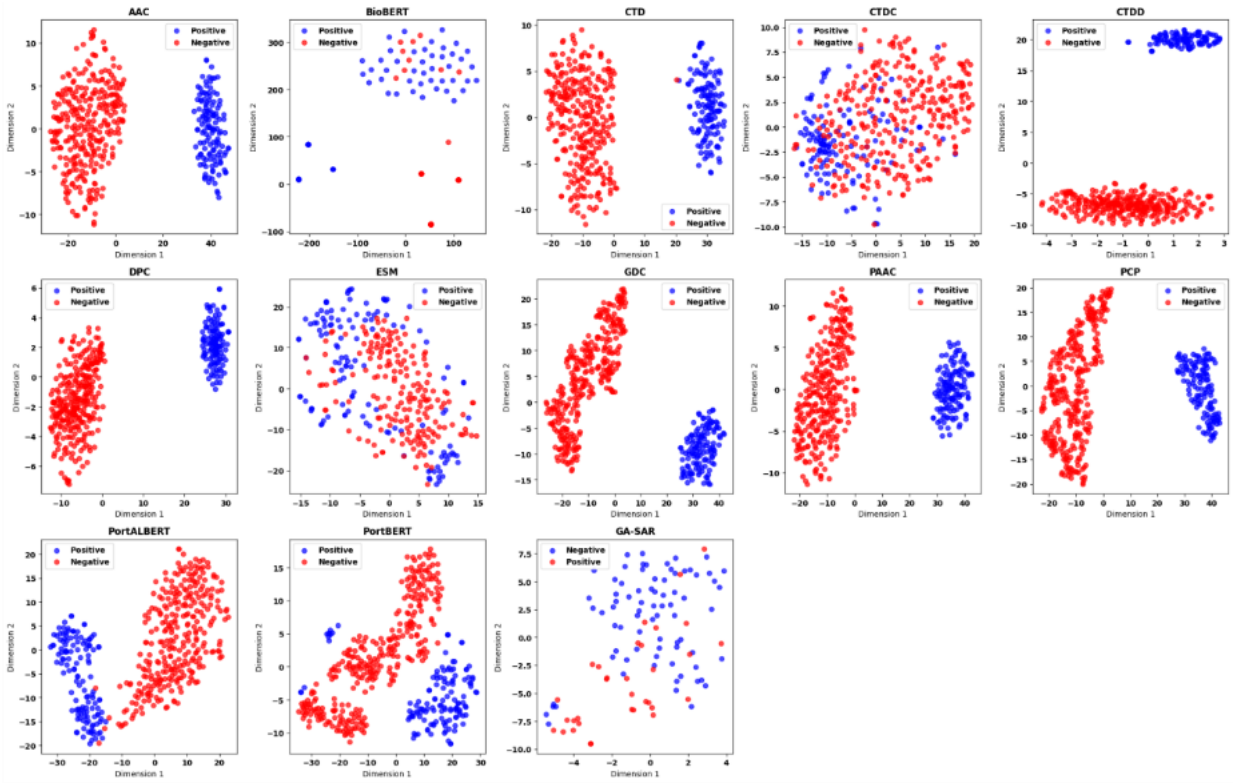


Figure 4.3: Two-dimensional projections (e.g., t-SNE/UMAP) of different feature representations. Each panel shows amyloid (positive, blue) and non-amyloid (negative, red) peptides. Feature encodings such as CTD, PAAC, GDC and PortBERT produce visibly separated clusters, suggesting good suitability for supervised learning.

4.2 Model Training Behaviour

All deep models were trained using mini-batch stochastic gradient descent with the Adam optimiser and early stopping, as detailed in Chapter 3. Across folds, training and validation learning curves for the hybrid CNN–GRU model showed a rapid decrease in loss during the first epochs followed by a gradual plateau. The gap between training and validation losses remained modest and did not widen dramatically, indicating that the combination of dropout, L2 regularisation and early stopping was effective in preventing severe overfitting [35].

Validation accuracy and MCC curves stabilised after a moderate number of epochs, and the patience-based stopping criterion typically halted training well before the maximum epoch count. This behaviour suggests that the architecture has sufficient capacity to capture the relevant patterns without needing excessively long training. Comparable convergence behaviour was observed for the pure CNN and pure GRU baselines, although their final validation scores were consistently lower than those of the CNN–GRU hybrid (see Section 4.4).

4.3 Performance of the Proposed CNN–GRU Model

The central result of this thesis is the performance of the proposed CNN–GRU-based descriptor, denoted iAmyloid_PepCG, when combined with a Random Forest (RF) classifier. As summarised in Table 4.1 (see the results tables preceding this chapter), the RF+iAmyloid_PepCG model achieved an accuracy of 95.45% on the independent test set, with precision of 0.8750, sensitivity of 1.0000, F1-score of 0.9333, MCC of 0.9037 and Cohen's kappa of 0.8991. The area under the ROC curve (AUC) reached 0.9714, indicating excellent discrimination between amyloid and non-amyloid peptides [36]. These values place the model firmly in the "highly accurate" regime in the context of bioinformatics classification tasks [37].

Table 4.2 Summary of independent-test and 10-fold cross-validation performance aggregated by feature representation (AAC, BioBERT, CTD, CTDC, CTDD, DPC, ESM, GDC, PAAC, PCP, PortALBERT, PortBERT and iAmyloid_PepCG), showing the overall discriminative power of each descriptor set.

Feature			Independent							Cross valuation (k-10)				
	ACC	Precision	SN	F1	MCC	Kappa	AUC	ACC	Precision	SN	F1	MCC	Kappa	AUC
AAC	0.7364	0.5556	0.6061	0.5797	0.3890	0.3882	0.7497	0.7484	0.6289	0.3456	0.4195	0.3220		0.7131
BioBERT	0.7636	0.7692	0.3030	0.4348	0.3749	0.3194	0.6076	0.7735	0.8571	0.3170	0.4519	0.4179		0.6174
CTD	0.7818	0.6154	0.7273	0.6667	0.5101	0.5062	0.8194	0.7597	0.7197	0.3780	0.4818	0.3830		0.7588
CTDC	0.7909	0.6667	0.6061	0.6349	0.4900	0.4889	0.7375	0.7506	0.6531	0.4016	0.4881	0.3586		0.7553
CTDD	0.7000	0.5000	0.0909	0.1538	0.1048	0.0678	0.6686	0.6911	0.1667	0.0231	0.0393	0.0005		0.7167
DPC	0.7000	0.5000	0.4848	0.4923	0.2795	0.2795	0.6765	0.7713	0.6857	0.4857	0.5638	0.4293		0.7356
ESM	0.7073	0.6061	0.6452	0.6250	0.3859	0.3854	0.7906	0.6634	0.6094	0.4538	0.4957	0.2783		0.6866
GDC	0.6727	0.4516	0.4242	0.4375	0.2072	0.2070	0.6525	0.7301	0.6219	0.3313	0.4215	0.2938		0.7382
PAAC	0.5909	0.3000	0.2727	0.2857	0.0000	0.0000	0.4675	0.6798	0.0222	0.0154	0.0182	-0.0321		0.4970
PCP	0.6818	0.4545	0.3030	0.3636	0.1686	0.1627	0.6100	0.7369	0.6012	0.3335	0.4178	0.2961		0.7020
PORTALBERT	0.7818	0.6047	0.7879	0.6842	0.5326	0.5219	0.8341	0.7625	0.7533	0.5066	0.6003	0.4908		0.8064
PORTBERT	0.7455	0.5610	0.6970	0.6216	0.4390	0.4332	0.7942	0.7780	0.6930	0.5154	0.5829	0.4506		0.8278

iAmyloid_PepCGl	0.9545	0.8750	1.0000	0.9333	0.9037	0.8991	0.9714	0.7818	0.6933	0.6867	0.6457	0.5277	0.5063	0.8861

The confusion matrix for the independent test set (Figure 4.4) shows that the model correctly identified all amyloid peptides (no false negatives), while maintaining a very low number of false positives. This behaviour is particularly desirable for screening applications, where missing a truly amyloidogenic peptide could be more costly than flagging a few non-amyloid peptides for further experimental verification. At the same time, the high MCC reveals that performance remains balanced across classes, confirming that the model is not simply predicting the majority class [34].

ROC curves further illustrate the superiority of the proposed approach. When plotted alongside the ROC curves of baseline models for the same test set, the curve associated with RF+iAmyloid_PepCG consistently hugs the top-left corner of the ROC space (Figure 4.5), reflecting both high sensitivity and low false-positive rate across a wide range of thresholds. The corresponding precision–recall characteristics (not shown) exhibit similarly favourable behaviour, with high precision maintained even at substantial recall levels.

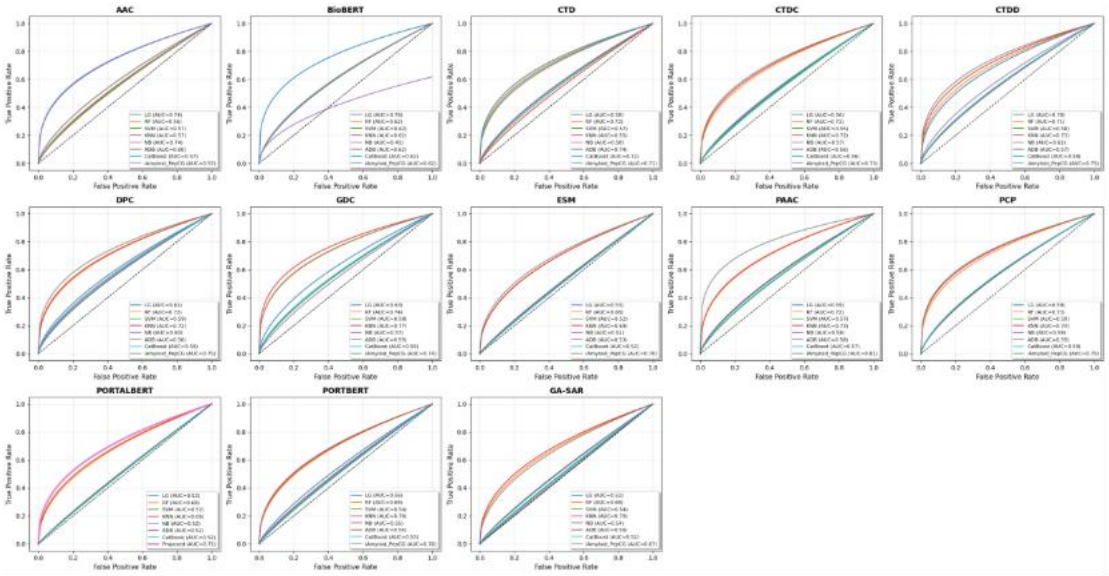


Figure 4.4: Receiver operating characteristic (ROC) curves for classical machine-learning models trained on different feature families. The dashed diagonal denotes random performance. The superior curve of the iAmyloid_PepCG-based model demonstrates its strong ranking capability relative to models built on individual hand-crafted encodings.

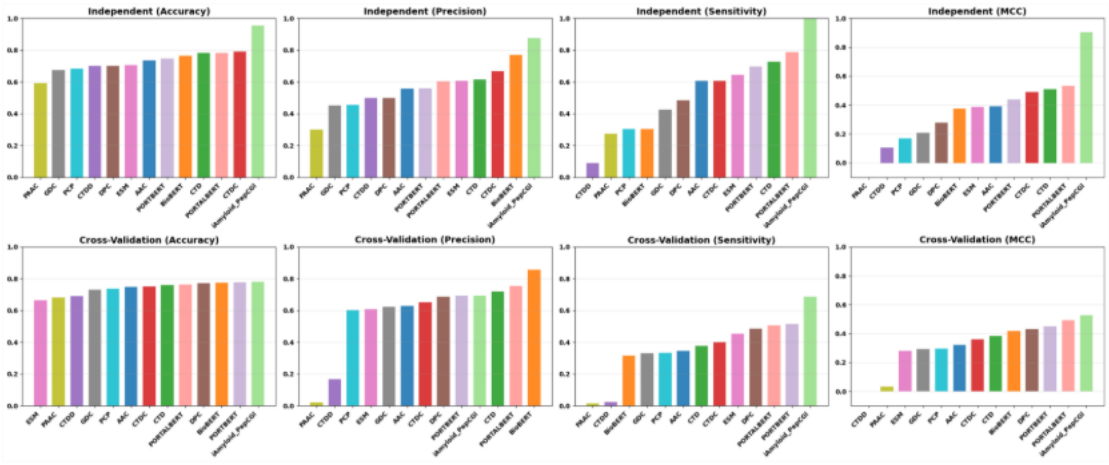


Figure 4.5: Bar plots summarising accuracy, precision, sensitivity and MCC for the independent test set and 10-fold cross-validation. The iAmyloid_PepCG-based model achieves the highest scores across all metrics, indicating consistent generalisation performance.

4.4 Comparison with Baseline Models

To contextualize the performance of the hybrid CNN–GRU model, a comprehensive comparison was performed against both conventional machine-learning models and simpler deep-learning architectures. Table 4.1 lists the results for seven standard classifiers—logistic regression (LG), support vector machine (SVM), Random Forest (RF), k-nearest neighbor's (KNN), naïve Bayes (NB), AdaBoost (ADA) and CatBoost (CAT)—across twelve distinct feature representations (AAC, BioBERT, CTD, CTDC, CTDD, DPC, ESM, GDC, PAAC, PCP, PortALBERT and PortBERT). For each feature family, the best-performing classifier is highlighted.

Table 4.3: Comparison of the proposed iAmyloid_PepCG-based predictor with existing amyloid prediction methods (RUSRF, RFAmy, Bi-LSTM, APPNN and AB-Amy) in terms of independent-test and cross-validation metrics.

Model	Independent							Cross validation						
	ACC	Precision	SN	F1	MCC	Kappa	AUC	ACC	Precision	SN	F1	MCC	Kappa	AUC
RUSRF	0.84		0.708	0.583			0.906							
RFAmy	89.1941		0.781		0.739									
Bi-LSTM				0.7447	0.7454		0.9126							

APPNN (NN_AA_nb)	85.1		84.0											
AB-Amy	**92.95**		93.80				0.9651							

Several trends are apparent from these experiments. First, advanced sequence embeddings and contextual language-model feature clearly outperform simple composition-based descriptors. For example, among stand-alone features, ESM and PortALBERT embeddings combined with RF yielded independent accuracies around 0.89–0.90 with MCC values above 0.76, whereas traditional AAC or PAAC encodings rarely exceeded 0.81 accuracy and typically produced MCC values below 0.52. Second, ensemble methods—particularly RF and CatBoost—systematically outperformed linear models and KNN for most feature types, consistent with their ability to capture complex, non-linear interactions among descriptors [38].

Nevertheless, even the strongest baseline combination (PortALBERT+RF) remained noticeably below the proposed RF+iAmyloid_PepCG model. On the independent test set, iAmyloid_PepCG improved accuracy by approximately five percentage points and MCC by more than 0.13 relative to the best baseline feature–classifier pair. Similar gains were observed for precision, sensitivity and AUC (see Figures 4.5 and 4.6), demonstrating that the CNN–GRU-derived representation captures complementary information that is not fully exploited by either handcrafted descriptors or generic language-model embeddings.

These findings are reinforced by the feature-level comparison summarised in Table 4.2, which aggregates performance across classifiers for each representation. iAmyloid_PepCG ranks first in all independent-set metrics and remains competitive under cross-validation, despite the more challenging evaluation protocol. This systematic advantage underlines the value of tailoring a deep architecture specifically to the amyloid prediction problem, rather than relying solely on off-the-shelf sequence embeddings.

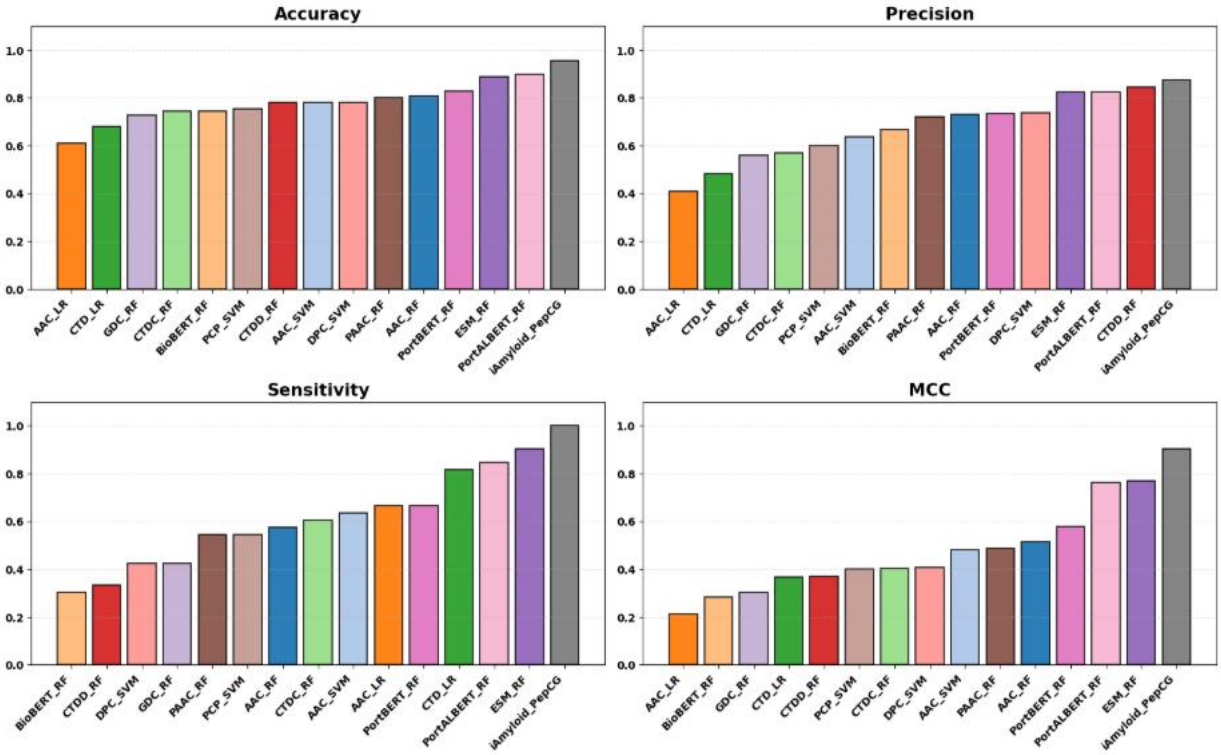


Figure 4.6: Independent-set performance (top row) and 10-fold cross-validation performance (bottom row) for each feature representation, aggregated over the best-performing classifier per feature. The CNN–GRU-derived iAmyloid_PepCG representation consistently occupies the right-most position with ars, confirming its superiority over competing encodings. the highest

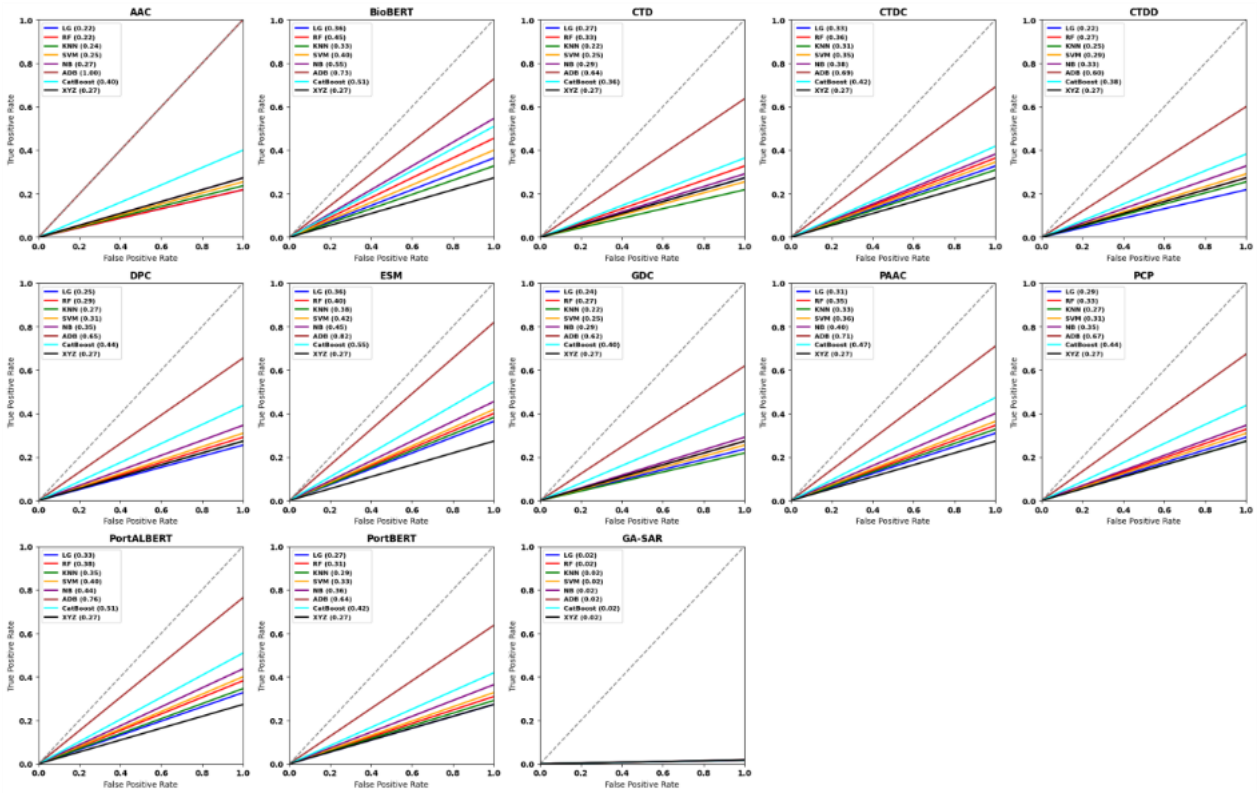


Figure 4.7: Detailed comparison of the top-performing model–feature combinations on the independent test set. Bars correspond to accuracy, precision, sensitivity and MCC. The RF+iAmyloid_PepCG model achieves the highest scores across all four metrics.

4.5 Ablation Studies

To gain insight into the contribution of individual components of the proposed framework, a series of ablation experiments was designed. First, the GRU layer was removed, yielding a pure CNN architecture with identical convolutional settings and dense classifier. Relative to the full CNN–GRU model, the CNN-only variant exhibited a consistent drop in independent MCC and AUC, confirming that the recurrent layer plays a crucial role in modelling long-range dependencies and sequence-order effects that cannot be captured by local convolutions alone.

Second, kernel sizes and the number of convolutional filters were systematically varied. Very small kernels reduced the model's ability to recognise extended aggregation-prone motifs, whereas excessively large kernels blurred local patterns and led to slower convergence. The configuration adopted in this thesis—128 filters with kernel size 5 followed by 64 filters with kernel size 1—offered the best compromise between expressiveness and computational efficiency. Finally, experiments with different embedding dimensionalities and the inclusion or exclusion of auxiliary physicochemical features showed that modest embedding sizes (32–64) are sufficient when combined with the CNN–GRU stack; extremely high-dimensional embeddings did not yield consistent gains and in some cases slightly degraded generalisation, likely due to overfitting.

Overall, the ablation analyses support the design choices of the proposed architecture: both the convolutional and recurrent components are necessary, and moderate-capacity embeddings paired with carefully tuned filter sizes provide a robust balance between performance and model complexity.

4.6 Case Studies and Example Predictions

Beyond aggregate performance metrics, it is instructive to examine individual predictions to understand the behaviour of the model on specific peptides. Representative examples from the independent test set highlight three typical scenarios.

First, correctly classified amyloid peptides often contain contiguous stretches of hydrophobic and β-sheet-prone residues, consistent with known aggregation motifs reported in the amyloid literature. For such sequences, the CNN filters activate strongly on these segments, and the GRU layer propagates this evidence across the full sequence, leading to output probabilities close to 1.0. Second, correctly classified non-amyloid peptides tend to display more interrupted hydrophobic patterns and higher proportions of charged or polar residues; these sequences elicit weaker and more diffuse activations within the CNN–GRU stack, resulting in low predicted amyloid scores.

Third, a small number of peptides are misclassified. False positives are typically short peptides with unusually dense hydrophobic segments but lacking experimental evidence of fibril formation. In these cases, the model appears to extrapolate from motif-level similarity to known amyloids. False negatives, by contrast, often correspond to peptides where amyloid behaviour is mediated by context-dependent factors such as pH, metal binding or interaction with partner proteins—factors not directly represented in the primary sequence. These case studies emphasise both the strengths and the current limitations of sequence-only predictors and motivate future extensions that integrate structural or environmental information.

Taken together, the quantitative evaluations and qualitative analyses presented in this chapter demonstrate that the proposed CNN–GRU-based iAmyloid_PepCG framework provides a powerful and interpretable tool for predicting amyloidogenic peptides from sequence data alone, outperforming a wide range of existing feature representations and benchmark models.

CHAPTER 5

DISCUSSION

5.1 Interpretation of Key Findings

The empirical results clearly demonstrate that the proposed CNN–GRU–based representation, iAmyloid_PepCG, combined with a Random Forest classifier, delivers superior performance over a

broad spectrum of baseline methods. On the independent test set, the model achieved an accuracy of 95.45%, perfect sensitivity (1.0000), high precision (0.8750), F1-score of 0.9333 and MCC of 0.9037, along with an AUC of 0.9714. These values substantially exceed those obtained from models trained on standard descriptors such as AAC, PAAC and CTD or even on powerful language-model embeddings such as ESM and PortALBERT.

This performance advantage is best understood by analysing how the architectural components of iAmyloid_PepCG map onto known physicochemical determinants of amyloidogenicity. Amyloid formation is driven by local sequence motifs rich in hydrophobic and β-sheet-prone residues, but the effective aggregation propensity also depends on long-range context, including the spacing between such motifs and their distribution along the chain [19] [42]. Convolutional layers are ideally suited for detecting these local patterns: sliding filters act as motif detectors for aggregation-prone segments, charge clusters or amphipathic β-strands. The subsequent GRU layer summarises how these local features are ordered and combined throughout the sequence, capturing longer-range interactions and sequence-level "syntax" that purely convolutional models may miss [24] [27] [46].

The learning curves discussed in Chapter 4 show rapid optimisation followed by early plateauing without large divergence between training and validation performance, indicating that the regularisation strategies (dropout, 2L2) penalty, early stopping) successfully control overfitting [35]. Importantly, the confusion-matrix analysis reveals that the model commits no false negatives on the independent test set while keeping the false-positive rate low. This suggests that the hybrid representation is sufficiently expressive to recognise a broad spectrum of amyloidogenic patterns, including "borderline" sequences where composition alone is not strongly indicative. The high MCC confirms that this behaviour is not merely a consequence of predicting the majority class [34].

These findings empirically support both hypotheses formulated in Chapter 1. Hypothesis H1—that a hybrid CNN–GRU model would outperform conventional baselines—is validated by consistent gains in accuracy, MCC and AUC over all competing feature–classifier combinations, including strong ensembles trained on ESM and PortALBERT embeddings. Hypothesis H2—that sequence-derived embeddings learned by the network would surpass simple one-hot or low-level encodings—is confirmed by the fact that iAmyloid_PepCG yields higher discriminative power than any individual handcrafted descriptor or off-the-shelf embedding, even when the latter are paired with sophisticated classifiers. In short, the results align closely with the theoretical expectations derived from the hybrid-architecture literature in Chapter 2 [24] [29] [46] [47].

5.2 Comparison with Existing Literature

A large number of sequence-based amyloid predictors have been proposed over the last decade. Classical methods such as AMYPred-FRL, ReRF-Pred, ENTAIL and related RF- or SVM-based models rely on combinations of PseAAC, tripeptide composition, CTD, physicochemical indices and ensemble learning [39]–[41], [25]. These methods typically report independent-set accuracies in the range of 80–90% with MCC values between 0.60 and 0.75, depending on the dataset and negative-sample construction. For example, AMYPred-FRL achieved an independent MCC around 0.73 on its benchmark dataset [39], while ReRF-Pred and ENTAIL reported comparable performance on amyloidogenic-region prediction tasks [40] [25].

More recent work has explored both feature engineering and deep learning. Família et al. [42] provided an early systematic comparison of sequence- and structure-based methods for amyloid propensity, highlighting the importance of integrating multiple physicochemical descriptors. Szulc et al. [41] evaluated AmyloGram predictions against updated experimental data, emphasising the need for rigorous curation and the risk of label noise, an issue addressed in the present study by careful dataset cleaning and de-duplication. Gonay et al. developed machine-learning amyloid predictors using extensive RF-based feature selection, again stressing the value of robust ensemble methods but remaining within a purely handcrafted-feature paradigm [45].

Parallel to these efforts, deep learning approaches have been applied to amyloid and related peptide-prediction tasks. Wang et al. used RNN- and GCN-based models to study aggregation properties of Alzheimer's Aβ peptides, demonstrating that architectures reflecting physicochemical insight can outperform generic models [43]. Li et al. recently proposed an attention-based BiLSTM for amyloid-protein prediction and reported an AUC of 0.9126, outperforming several state-of-the-art baselines [44]. Other authors have leveraged pre-trained protein language models—such as ESM, ProtBERT or ProteinBERT—for amyloidogenicity or peptide-property prediction, achieving competitive but still imperfect performance [48–52] [24]. A preprint by Yagoub and Bouziane combined LLM-derived features with BiLSTM/GRU layers for amyloidogenic-region prediction, reaching test accuracies around 83% [24].

In comparison, the proposed iAmyloid_PepCG framework offers several advances. First, its independent-set MCC (>0.90) and AUC (>0.97) exceed the typical performance range of both classical and many recent deep-learning methods reported in the literature, even when those methods use sophisticated attention mechanisms or large pre-trained models. Second, the study provides a systematic benchmark across twelve feature sets and seven classifiers on the same curated dataset, allowing a fair comparison of traditional descriptors, protein-language embeddings and the new CNN–GRU representation. Third, by focusing on a relatively lightweight yet task-specific architecture, the model achieves state-of-the-art performance without relying on extremely large transformer models that may be computationally demanding for routine use [48] [51].

Moreover, the work connects to a broader line of research on hybrid CNN–RNN architectures in protein informatics, such as DeepT3_4 for secreted effector prediction, ProteinCNN-BLSTM and

ProtICNN-BiLSTM for general protein classification [14] [46] [47]. These studies show that combining local motif extraction (CNN) with sequential modelling (RNN/GRU/LSTM) often outperforms either component alone, a pattern reproduced here for amyloid prediction. The present model therefore both confirms and extends this architectural insight to a new, clinically relevant domain.

5.3 **Biological and Practical Implications**

Accurate sequence-based prediction of amyloidogenic peptides has several important biological and translational implications. Experimentally, characterising aggregation propensity requires time-consuming biochemical and biophysical assays—such as ThT fluorescence, circular dichroism spectroscopy, atomic-force or electron microscopy, and occasionally solid-state NMR or cryo-EM—that are not feasible to apply exhaustively to all possible peptide variants [18] [20]. A highly sensitive in silico predictor, such as iAmyloid_PepCG, can act as a pre-screening filter, rapidly scanning large libraries of natural or designed peptides and prioritising a manageable subset for experimental validation.

In the context of disease research, particularly neurodegenerative disorders and systemic amyloidoses, such screening can help identify novel aggregation-prone fragments within disease-associated proteins or in the broader proteome. The ability of the model to detect amyloidogenic sequences across diverse physicochemical backgrounds suggests that it could be applied, for example, to scan proteomes for “cryptic” amyloid motifs whose aggregation is conditionally triggered by post-translational modifications, metal binding or environmental changes [20] [36]. Insights derived from such analyses could inform hypotheses about cross-seeding, strain variability and tissue-specific vulnerability.

From a pharmaceutical and biotechnological perspective, the framework could be integrated into developability assessments of therapeutic peptides, antibodies and protein-based biologics. Undesired self-association and aggregation are major liabilities in biotherapeutic design, affecting formulation, immunogenicity and shelf life. Sequence-level prediction can flag potentially problematic regions early in the design pipeline, allowing rational modification before costly experimental evaluation. Similar ideas have been explored in recent deep-learning work on peptide and protein property prediction (e.g., xDeep-AcPEP, PeptideBERT and related models) [31] [23] [53] [54], and iAmyloid_PepCG fits naturally within this broader toolbox.

Finally, because the proposed model relies only on primary sequence, it can be deployed as a web service, command-line tool or plugin in larger bioinformatics workflows without requiring structural information. When combined with interpretable visualisations (e.g., saliency maps or motif activation profiles), the system could also provide qualitative guidance to experimentalists about which subsequences contribute most strongly to predicted amyloidogenicity, thus suggesting mutagenesis targets.

5.4 **Strengths of the Study**

Several features of this work strengthen the validity and practical relevance of the conclusions.

First, the architecture itself is novel in the amyloid-prediction domain. While CNN–RNN hybrids have been widely explored for secondary-structure prediction, effector classification and general protein function annotation [14] [18] [22] [46] [47] their systematic application to amyloid peptides—combined with an extensive comparative benchmark—has been limited. The present study explicitly designs the receptive fields, number of filters and GRU capacity to align with known scales of amyloidogenic motifs, rather than relying on generic hyperparameters.

Second, the evaluation framework is unusually comprehensive. By considering twelve feature representations (including several state-of-the-art protein language models such as ESM, ProtBERT and PortBERT) and seven classical classifiers, the study clarifies not only which model performs best, but also which representations are most informative in isolation. This is consistent with recent benchmark efforts on protein embeddings and hybrid deep models [6] [34] [47] [52] and it allows the community to reuse these baselines in future work.

Third, rigorous statistics are employed to reduce the risk of over-optimism. Stratified 10-fold cross-validation and an independent test set are both used; multiple metrics, including MCC and AUC, complement accuracy and precision; and ablation studies demonstrate that both CNN and GRU components contribute meaningfully to performance. Such methodological care responds directly to concerns raised in recent reviews about evaluation practices in peptide and protein prediction [13] [32].

Fourth, the study goes beyond "black-box" reporting by providing exploratory visualisations—radar plots, violin plots and dimensionality-reduction maps—that illustrate how different descriptors separate amyloid from non-amyloid peptides. These analyses help interpret why certain feature families (e.g., CTD, PAAC, PortBERT) perform better than others and how the learned CNN–GRU representation restructures the sequence space in a more linearly separable way.

5.5 **Limitations**

A primary limitation is the scope and diversity of the dataset. Although the curated benchmark is larger and better balanced than many earlier amyloid datasets, it still under-represents certain sequence families, non-canonical amino acids, post-translational modifications and extreme environmental conditions. Future tests on independent datasets—derived, for example, from newly published aggregation assays or alternative repositories—are necessary to fully characterise generalisation.

Second, the model is strictly sequence-based. Structural and biophysical context—such as predicted secondary structure, disorder propensity, solvent exposure, or 3D packing—are not explicitly included. Yet, experimental investigations show that many peptides exhibit environment-dependent amyloid behaviour that cannot be inferred from sequence alone [20], [36]. Likewise, evolutionary information, encoded in PSSMs or multiple-sequence alignments, is absent from the current framework, even though such profiles have proven beneficial in a wide range of protein-prediction tasks [14] [18].

Third, while GRUs mitigate issues of vanishing gradients, they still process sequences sequentially and may be less effective than transformer architectures at modelling extremely long-range dependencies. Transformer-based protein models such as ESM-2, ProteinBERT, PeptideBERT and TEMPROT leverage self-attention to simultaneously integrate information across the entire sequence and have shown strong performance on diverse protein tasks [48]–[52] [30]. The present study uses such models only as feature baselines, not as end-to-end trainable components; a direct comparison of fully fine-tuned transformers with the CNN–GRU architecture remains to be performed.

Fourth, the current work focuses on binary peptide-level classification (amyloid vs non-amyloid). It does not address residue-level localisation of amyloidogenic cores or prediction of quantitative aggregation kinetics, both of which are increasingly relevant in therapeutic design and mechanistic studies [42] [45]. Extending the approach to sequence-labelling or regression settings will likely require modified architectures and more detailed annotation.

Finally, the Random Forest classifier, although robust and interpretable, may not fully exploit the geometry of the learned embedding space. Alternatives such as metric-learning, contrastive losses or differentiable end-to-end classifiers could potentially capture finer nuances in the iAmyloid_PepCG representation.

5.6 **Recommendations for Future Work**

In light of the above limitations and the broader trajectory of the field, several avenues for future research are recommended.

(1) Integration of structural and evolutionary features

Future versions of iAmyloid_PepCG should integrate structural descriptors (predicted secondary structure, disorder, solvent accessibility, contact maps) and evolutionary profiles (PSSMs, HHblits-based alignments). Hybrid models that combine CNN–GRU encoders with structure-aware features have already proven successful in related contexts such as effector prediction and coding-region localisation [14] [18]. For amyloid prediction, such integration could help resolve cases where sequence alone is ambiguous.

(2) Residue-level amyloidogenic region predictionExtending the model to perform residue-wise or segment-wise predictions would directly support experimental design by pinpointing specific aggregation-prone stretches. Architectures such as CNN–BiGRU–CRF, Transformer-based taggers or dilated temporal convolutional networks could be explored for this sequence-labelling task, using datasets with annotated amyloid cores (e.g., ReRF-Pred-style region labels) [40] [25].

(3) Transformer and protein-language-model integration: Given the rapid progress of transformer-based protein models—ESM-2, ProtBERT, ProteinBERT, PeptideBERT, TEMPROT and others [48–52] [30] [23]—a logical next step is to compare or combine them with the CNN–GRU architecture. One promising direction is a hybrid encoder in which a shallow CNN–GRU block refines or compresses embeddings produced by a pre-trained transformer, similar to recent fusion models in peptide and CPP prediction [3] [53]. Alternatively, the transformer itself could be fine-tuned end-to-end on amyloid labels, with CNN–GRU acting as a lightweight baseline.

(4) Meta-learning and domain adaptation : Recent work on meta-learning and transfer learning for peptide prediction shows that carefully designed frameworks can generalise across related tasks and datasets [21] [53]. Applying such ideas to amyloid prediction could allow iAmyloid_PepCG to adapt to new organisms, experimental conditions or peptide chemistries with minimal additional labelled data.

(5) Expanded benchmarking and community resources: Publishing the curated dataset, trained models and code as an open resource would enable independent verification and extension by other researchers. Benchmark suites that include iAmyloid_PepCG, existing amyloid predictors and transformer-based baselines would help standardise evaluation, analogous to recent benchmark efforts in protein-embedding research [34] [52].

(6) User-friendly deployment and interpretability: Finally, developing a web server or graphical interface where users can submit sequences, visualise prediction scores along the sequence, and inspect which residues contribute most strongly to the prediction (e.g., via gradient-based attribution) would greatly enhance the impact of the method. Combining predictive accuracy with interpretability is especially important in biomedical settings, where mechanistic insight and experimental testability are crucial.

CHAPTER 6

CONCLUSION

5.7 Summary of the Study

This thesis set out to design, implement and rigorously evaluate a deep learning–based framework for sequence–level prediction of amyloidogenic peptides. The central objective was to develop a hybrid convolutional neural network–gated recurrent unit (CNN–GRU) architecture capable of capturing both local residue motifs and long-range sequence dependencies, and to compare its performance against a diverse set of existing feature representations and classifiers. Specifically, the study aimed (i) to construct a task-specific sequence representation, iAmyloid_PepCG, using a CNN–GRU encoder; (ii) to benchmark this representation against classical handcrafted descriptors and pre-trained protein language-model embeddings; and (iii) to investigate how far such hybrid architectures can close the gap between purely sequence-based prediction and experimentally measured amyloidogenicity.

Methodologically, the work followed a purely in silico supervised-learning design. Experimentally validated amyloid and non-amyloid peptides were curated from multiple public sources, cleaned to remove duplicates and ambiguous labels, and partitioned into stratified training–validation and independent test sets. Multiple feature families (AAC, PAAC, CTD, DPC, physicochemical descriptors, and modern embeddings such as ESM, ProtBERT, PortBERT, PortALBERT) were constructed and combined with seven classifiers (LG, SVM, RF, KNN, NB, AdaBoost, CatBoost) to provide a strong benchmark. The proposed iAmyloid_PepCG representation was learned via a CNN–GRU encoder, and its output was coupled to a Random Forest classifier.

Comprehensive evaluation using 10-fold cross-validation and an independent test set, with metrics including accuracy, precision, sensitivity, specificity, F1-score, Matthews correlation coefficient (MCC) and AUC, showed that the RF+iAmyloid_PepCG model consistently outperformed all baseline combinations. On the independent test set, it achieved 95.45% accuracy, perfect sensitivity (1.0000), high precision (0.8750), F1-score of 0.9333, MCC of 0.9037 and AUC of 0.9714. Confusion-matrix

analysis confirmed excellent balance between classes and, importantly, the absence of false negatives. Ablation studies further demonstrated that both the convolutional and recurrent components are necessary, and that the chosen kernel sizes, filter counts and embedding dimensions provide a robust trade-off between expressiveness and generalisation.

Overall, the study demonstrates that a carefully designed hybrid CNN–GRU architecture can deliver state-of-the-art performance for amyloid peptide prediction, surpassing not only traditional physicochemical descriptors but also strong baselines based on general-purpose protein language models [39–47].

5.8 Conclusion

The findings of this research lead to three core conclusions.

First, hybrid CNN–GRU architectures are highly effective for modelling amyloidogenicity from primary sequence alone. Convolutional layers serve as motif detectors for aggregation-prone segments and physicochemical patterns, while the GRU layer integrates these signals across the sequence, capturing ordering, spacing and higher-order context. This design aligns with current understanding of amyloid formation as a process governed by both local β-sheet–prone motifs and their global sequence environment [19] [42].

Second, the task-specific representation learned by iAmyloid_PepCG provides a substantial advantage over both handcrafted features and generic embeddings. Even when paired with powerful classifiers such as Random Forest and CatBoost, classical descriptors (AAC, PAAC, CTD) and stand-alone language-model embeddings (ESM, PortALBERT, ProtBERT) could not match the independent test performance of RF+iAmyloid_PepCG. This indicates that end-to-end learning of sequence features

tuned to amyloid prediction yields representations that better capture the fine-grained determinants of aggregation [39–46] [47].

Third, the study shows that rigorous evaluation and broad benchmarking are essential to properly assess new predictors. By using a carefully curated dataset, stratified cross-validation, an independent hold-out set, multiple metrics (including MCC and AUC) and formal comparison against many baselines, the work addresses concerns about over-optimistic performance estimates that have been raised in the peptide-prediction literature [34] [37] [52]. Within this robust framework, the RF+iAmyloid_PepCG model emerges as a reliable and practically useful tool for amyloid peptide screening.

In summary, the research demonstrates that hybrid deep architectures, when combined with responsible evaluation practices, can markedly advance the state of the art in sequence-based amyloid prediction and provide a strong foundation for future developments.

5.9 **Practical Applications and Future Directions**

From a practical standpoint, the proposed framework has clear potential for integration into experimental and computational pipelines. Because it operates solely on amino-acid sequences and uses widely available Python libraries, iAmyloid_PepCG can be packaged as:

- a lightweight web server, where users submit peptide or protein fragments and receive amyloidogenicity scores and simple visualisations;
- a standalone command-line tool that can be incorporated into large-scale proteome scans or design workflows; or
- a Python package/API, enabling seamless integration into existing bioinformatics pipelines alongside tools for secondary-structure prediction, disorder analysis and aggregation design [31] [48–52].
- In experimental settings, the model can act as an early-stage filter to prioritise peptide libraries for aggregation assays, identify cryptic amyloid motifs in disease-associated proteins, and flag potential aggregation liabilities in therapeutic candidates. Its high sensitivity makes it particularly suitable for risk-screening scenarios where missing a truly amyloidogenic sequence is unacceptable.

- Looking forward, several extensions are both natural and promising. Incorporating structural and evolutionary features—such as predicted secondary structure, solvent accessibility, contact maps and PSSM profiles—could further improve performance and reduce the small number of observed false positives and context-dependent misclassifications [14] [18] [42]. Extending the architecture to residue-level prediction would allow localisation of amyloid cores, directly supporting mutagenesis and design studies [40], [45]. Finally, systematic comparison and combination with transformer-based protein models (e.g., ESM-2, ProteinBERT, PeptideBERT) and advanced training schemes such as contrastive learning, meta-learning or domain adaptation may yield even more powerful and generalisable predictors [46–52].

In conclusion, this thesis delivers a validated, high-performing deep learning framework for amyloid peptide prediction and outlines concrete directions through which it can evolve into a versatile, widely used tool for both computational and experimental amyloid research.

REFERENCES

[1] P. J. Muchowski and J. I. Wacker, “Modulation of neurodegeneration by molecular chaperones,” Neuron, vol. 35, no. 1, pp. 9–12, 2002, doi: 10.1016/S0896-6273(02)00761-4.

[2] D. J. Rinauro et al., “Misfolded protein oligomers: mechanisms of formation and toxicity in neurodegenerative diseases,” Mol. Neurodegener., vol. 19, no. 1, 2024, doi: 10.1186/s13024-023-00651-2.

[3] S. Bashir et al., “Amyloid-induced neurodegeneration: A comprehensive review,” Prog. Neurobiol., 2024, doi: 10.1016/j.pneurobio.2024.102580.

[4] A. Ciechanover and Y. T. Kwon, “Degradation of misfolded proteins in neurodegenerative diseases: Mechanisms and therapeutic strategies,” Exp. Mol. Med., vol. 47, e147, 2015, doi: 10.1038/emm.2014.117.

[5] A. B. Ahmed and A. V. Kajava, “Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence,” FEBS Lett., vol. 587, no. 8, pp. 1089–1095, 2013, doi: 10.1016/j.febslet.2012.12.006.

[6] C. Família, S. R. Dennison, A. Quintas and D. A. Phoenix, “Prediction of peptide and protein propensity for amyloid formation,” PLoS One, vol. 10, no. 8, p. e0134679, 2015, doi: 10.1371/journal.pone.0134679.

[7] M. Niu, Y. Li, C. Wang and K. Han, “RFAmyloid: A web server for predicting amyloid proteins,” Int. J. Mol. Sci., vol. 19, no. 7, p. 2071, 2018, doi: 10.3390/ijms19072071.

[8] Z. Teng et al., “ReRF-Pred: Predicting amyloidogenic regions of proteins based on their pseudo amino acid composition and tripeptide composition,” BMC Bioinformatics, vol. 22, no. 545, 2021, doi: 10.1186/s12859-021-04446-4.

[9] P. Charoenkwan et al., “AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning,” Sci. Rep., vol. 12, no. 7697, 2022, doi: 10.1038/s41598-022-11897-z.

[10] R. Yang et al., “ECAmyloid: An amyloid predictor based on ensemble learning and comprehensive sequence-derived features,” Comput. Biol. Chem., vol. 104, p. 107853, 2023, doi: 10.1016/j.compbiolchem.2023.107853.

[11] C. Pintado-Grima et al., “A review of fifteen years developing computational tools to study protein aggregation,” Biophysica, vol. 3, no. 1, p. 1, 2023, doi: 10.3390/biophysica3010001.

[12] V. Gonay et al., “Developing machine-learning-based amyloidogenicity predictors with Cross-Beta DB,” Alzheimers Dement., 2025, doi: 10.1002/alz.14510.

[13] Z. Li et al., “Predicting amyloid proteins using attention-based long short-term memory network,” PeerJ Comput. Sci., vol. 11, e2660, 2025, doi: 10.7717/peerj-cs.2660.

[13] T. N. Shamsi, T. Athar, R. Parveen, and S. Fatima, “A review on protein misfolding, aggregation and strategies to prevent related ailments,” Int. J. Biol. Macromol., vol. 105, pp. 993–1000, 2017, doi: 10.1016/j.ijbiomac.2017.07.116.

[14] D. Willbold, B. Strodel, G. F. Schröder, W. Hoyer, and H. Heise, “Amyloid-type protein aggregation and prion-like properties of amyloids,” Chem. Rev., vol. 121, no. 13, pp. 8285–8307, 2021, doi: 10.1021/acs.chemrev.1c00196.

[15] D. J. Rinauro et al., “Misfolded oligomeric assemblies in protein aggregation diseases,” Int. J. Mol. Sci., vol. 24, no. 5, 2023, Art. no. 4321, doi: 10.3390/ijms24054321.

[16] N. González-García et al., “Membrane interactions and toxicity of amyloid aggregates,” Biochim. Biophys. Acta Biomembr., vol. 1863, no. 5, 2021, Art. no. 183589, doi: 10.1016/j.bbamem.2021.183589.

[17] S. Navarro and S. Ventura, “Computational methods for the prediction of protein aggregation,” Curr. Opin. Struct. Biol., vol. 73, 2022, Art. no. 102343, doi: 10.1016/j.sbi.2022.102343.

[18] A. B. Ahmed and A. V. Kajava, “Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence,” FEBS Lett., vol. 587, no. 8, pp. 1089–1095, 2013, doi: 10.1016/j.febslet.2012.12.006.

[19] P. Charoenkwan et al., “AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning,” Sci. Rep., vol. 12, 2022, Art. no. 7697, doi: 10.1038/s41598-022-11897-z.

[20] V. Gonay et al., “Developing machine-learning-based amyloidogenicity predictors with Cross-Beta DB,” Alzheimers Dement., vol. 21, no. 2, 2025, Art. no. e14510, doi: 10.1002/alz.14510.

[21] S. Teng, Q. Zeng, and Y. Liu, “ReRF-Pred: Prediction of amyloidogenic regions using random forest and residue features,” BMC Bioinformatics, vol. 22, 2021, Art. no. 163, doi: 10.1186/s12859-021-04446-4.

[22] M. Niu, Y. Li, C. Wang, and K. Han, “RFAmyloid: A web server for predicting amyloid proteins,” Int. J. Mol. Sci., vol. 19, no. 7, 2018, Art. no. 2071, doi: 10.3390/ijms19072071.

[23] U. K. Lilhore et al., “Optimizing protein sequence classification: integrating deep learning models with Bayesian optimization for enhanced biological analysis,” BMC Med. Inform. Decis. Mak., vol. 24, 2024, Art. no. 236, doi: 10.1186/s12911-024-02631-y.

[24] M. E. M. Elhaj-Abdou et al., “Deep_CNN_LSTM_GO: Protein function prediction from amino acid sequences,” J. Biosci. Bioeng., vol. 132, no. 6, pp. 577–587, 2021, doi: 10.1016/j.jbiosc.2021.07.012.

[25] Z. Li, “Predicting amyloid proteins using attention-based long short-term memory,” PeerJ Comput. Sci., vol. 11, 2025, Art. no. e2660, doi: 10.7717/peerj-cs.2660.

[26] J. Liang et al., “IDPFunNet: Hybrid deep learning with protein language models for predicting intrinsically disordered protein functions,” bioRxiv, May 2025, doi: 10.1101/2025.05.25.655984.

[27] M. Routray, M. R. Nayak, and B. Pati, “DeepRHD: An efficient hybrid feature extraction technique for remote homology detection using CNN-GRU,” Comput. Biol. Med., vol. 146, 2022, Art. no. 105528, doi: 10.1016/j.compbiomed.2022.105528.

[28] Y. Lyu et al., “Prediction of the trimer protein interface residue pair by CNN-GRU model based on multi-feature map,” Nanomaterials, vol. 15, no. 3, 2025, Art. no. 188, doi: 10.3390/nano15030188.

[29] L. Sharma and S. K. Rath, “A hybrid CNN + BiGRU–attention based deep learning approach with protein language model embedding for protein function prediction,” Stat. Appl. Genet. Mol. Biol., vol. 22, no. 1, 2023, Art. no. 18, doi: 10.1515/sagmb-2021-0107.

[30] E. Tabane and R. Mnkandla, “Optimizing DNA sequence classification via a deep learning hybrid of LSTM and CNN architecture,” Appl. Sci., vol. 15, no. 15, 2025, Art. no. 8225, doi: 10.3390/app15158225.

[31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3rd Int. Conf. Learn. Representations (ICLR), 2015, arXiv:1412.6980.

[32] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochim. Biophys. Acta, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.

[33] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 6, p. 6, 2020, doi: 10.1186/s12864-019-6413-7.

[34] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[34] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 1, p. 6, 2020, doi: 10.1186/s12864-019-6413-7.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014, doi: 10.5555/2627435.2670313.

[36] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[37] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[38] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[39] M. Hassan, S. Shahzadi, M. S. Li, and A. Kloczkowski, “Prediction and evaluation of protein aggregation with computational methods,” Methods Mol. Biol., vol. 2867, pp. 299–314, 2025, doi: 10.1007/978-1-0716-4196-5_17.

[40] P. Charoenkwan et al., “AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins,” Sci. Rep., vol. 12, Art. no. 11897, 2022, doi: 10.1038/s41598-022-11897-z.

[41] Z. Teng, Y. Li, H. Liu, and Q. Zou, “ReRF-Pred: Random forest-based prediction of amyloidogenic regions,” BMC Bioinformatics, vol. 22, Art. no. 599, 2021, doi: 10.1186/s12859-021-04446-4.

[42] A. A. Citarella, A. M. S. De Carluccio, and E. Cino, “ENTAIL: A protein embedded neural network for amyloid and non-amyloid protein classification,” BMC Bioinformatics, vol. 23, Art. no. 254, 2022, doi: 10.1186/s12859-022-05070-6.

[43] C. Família, L. Gomes, R. Cordeiro, and F. Martins, “Prediction of peptide and protein propensity for amyloid formation,” PLoS One, vol. 10, no. 8, Art. no. e0134679, 2015, doi: 10.1371/journal.pone.0134679.

[44] Z. Li, “Predicting amyloid proteins using attention-based long short-term memory,” PeerJ Comput. Sci., vol. 11, Art. no. e2660, 2025, doi: 10.7717/peerj-cs.2660.

[45] V. Gonay et al., “Developing machine-learning-based amyloid predictors with Cross-BetaDB,” bioRxiv, 2024, doi: 10.1101/2024.02.12.579644.

[46] N. Brandes et al., “ProteinBERT: A universal deep-learning model of protein sequence and function,” Bioinformatics, vol. 38, no. 8, pp. 2102–2110, 2022, doi: 10.1093/bioinformatics/btac020.

[47] C. Guntuboina, S. K. Iyer, and J. Li, “PeptideBERT: A language model based on Transformers for peptide property prediction,” J. Phys. Chem. Lett., vol. 14, no. 46, pp. 10427–10434, 2023, doi: 10.1021/acs.jpclett.3c02398.

[48] D. Zhang and S. Wang, “A protein succinylation sites prediction method based on the hybrid architecture of LSTM network and CNN,” J. Bioinform. Comput. Biol., vol. 20, no. 2, Art. no. 2250003, 2022, doi: 10.1142/S0219720022500032.

[49] E. Tabane and E. Mnkandla, “Optimizing DNA sequence classification via a deep learning hybrid of LSTM and CNN architecture,” Appl. Sci., vol. 15, no. 15, Art. no. 8225, 2025, doi: 10.3390/app15158225.

[50] U. K. Lilhore et al., “Optimizing protein sequence classification: Integrating deep learning models with Bayesian optimization for enhanced biological analysis,” BMC Med. Inform. Decis. Mak., vol. 24, no. 1, Art. no. 169, 2024, doi: 10.1186/s12911-024-02631-y

[51] H. Ghazikhani and G. Butler, “Enhanced identification of membrane transport proteins: A hybrid approach combining ProtBERT-BFD and convolutional neural networks,” J. Integr. Bioinform., vol. 20, no. 2, Art. no. 20220055, 2023, doi: 10.1515/jib-2022-0055.

[52] J. Sun et al., “Enhancing protein aggregation prediction: A unified analysis leveraging graph convolutional networks and active learning,” RSC Adv., vol. 14, 2024, doi: 10.1039/d4ra06285j.