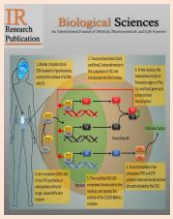
RESEARCH ARTICLE

Biological Sciences
(An International Journal of Medical, Pharmaceutical, and Life Sciences)

Journal home: https://irrespub.com/biolsciences/index.php/1/index



Comparative Analysis of Bagging and Boosting Algorithms for Predicting Protein-Protein Interactions Using Learned Embeddings

Sini S Raj*, Vinod Chandra S S

Machine Intelligence Research Lab, Department of Computer Science, University of Kerala, Thiruvananthapuram, Kerala, India.

ARTICLE HISTORY

Received: 19-01-2026
Revised: 22-02-2026
Accepted: 04-03-2026
Online: 13-03-2026

KEYWORDS

Viral infections
Protein-Protein Interactions (PPIs)
Word2Vec embedding
Learned embeddings
Ensemble learning

ABSTRACT

Viral infections are a major global health concern, as evidenced by the rapid spread of SARS-CoV-2, leading to a worldwide pandemic. Viruses can manipulate host cell machinery by integrating their genetic material into the host genome, a process facilitated by Protein-Protein Interactions (PPIs). Identifying PPIs between humans and viruses is essential for understanding the mode of infection and host immune responses and developing effective treatment regimes. Although experimental methods like mass spectrometry-based proteomics and yeast two-hybrid assays are widely employed to identify human-virus PPIs they are often time-consuming, expensive, and labor-intensive. Here, we propose an alternative method that overcomes technical limitations by leveraging machine learning models to predict human-virus PPIs with enhanced accuracy and efficiency, emphasizing the role of automatic feature extraction and ensemble learning techniques in driving superior prediction performance. Protein sequences are analyzed using Word2Vec embeddings to automatically extract complex features, offering a significant advantage over manual feature engineering. The study employs two ensemble learning approaches, boosting and bagging, to train predictive models on the extracted features. Among these, XGBoost, a boosting algorithm, demonstrated superior predictive performance compared to bagging models. Our findings highlight the potential of combining automated feature extraction with advanced ensemble learning methods to improve the efficiency and accuracy of PPI prediction. This approach enhances our understanding of protein sequences and their interactions and holds promise for accelerating the development of effective antiviral therapies.

*Address for correspondence

Machine Intelligence Research Lab, Department of Computer Science, University of Kerala, Thiruvananthapuram, Kerala, India.
Email: sinisraj@keralauniversity.ac.in; vinod@keralauniversity.ac.in
DOI: https://doi.org/10.55006/biolsciences.2026.6102
Published by IR Research Publication;

Introduction

Human–Virus Protein–Protein Interactions (PPIs) lie at the core of viral infection, determining how viruses attach to host receptors, enter cells, manipulate intracellular pathways, and evade immune surveillance. These interactions enable viruses to redirect essential host processes in order to replicate and persist. A comprehensive understanding of host–virus PPI networks is

therefore fundamental for elucidating infection mechanisms and identifying potential therapeutic targets. However, experimental approaches used to characterize these interactions are labor-intensive, costly, and often incomplete, particularly when studying newly emerging pathogens. These practical limitations highlight the need for reliable computational frameworks capable of predicting human–virus PPIs with both accuracy and biological relevance.

The urgency of developing such predictive frameworks has become increasingly evident in light of recent global viral outbreaks. The coronavirus disease 2019 (COVID-19) pandemic exemplifies the profound impact that emerging viral pathogens can have on public health and socioeconomic systems worldwide. Among these pathogens, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first identified in Wuhan, China, rapidly escalated into a global health crisis and was declared a pandemic by the World Health Organization in early 2020 (1–3). Earlier outbreaks caused by SARS-CoV in 2002 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 had already demonstrated the capacity of coronaviruses to cause severe disease (4,5). The emergence of SARS-CoV-2 further reinforced the recognition that coronaviruses belong to a broader class of zoonotic RNA viruses capable of crossing species barriers and establishing sustained transmission in human populations.

Importantly, this pattern of zoonotic emergence extends beyond coronaviruses. Several RNA viruses, including Ebola virus, Zika virus, and Nipah virus, have repeatedly spilled over from animal reservoirs and triggered outbreaks with significant regional and global consequences (6–8). Despite differences in transmission routes and clinical manifestations, these viruses share common biological characteristics, such as high mutation rates, rapid evolutionary adaptation, and sophisticated immune evasion strategies. These properties complicate vaccine and antiviral development and make it difficult to anticipate future outbreaks. Collectively, these recurring events underscore the need for predictive computational models capable of identifying interaction patterns between viral and host proteins across diverse viral families.

At the molecular level, viral pathogenicity is fundamentally driven by interactions between viral and host proteins (9–11). Through these interactions, viruses modulate cellular signaling pathways, alter immune responses, and reprogram host metabolic processes to establish infection. Mapping these molecular interactions is therefore essential for understanding the mechanisms of viral replication and host adaptation. However, the scale and complexity of host–virus interactomes make comprehensive experimental mapping impractical, particularly for newly emerging viruses. Computational prediction has thus become an indispensable tool for large-scale identification of potential PPIs.

Despite substantial advances in computational biology, current sequence-based PPI prediction methods face notable challenges. Many approaches rely on handcrafted sequence descriptors or limited feature representations that may fail to capture contextual relationships within protein sequences, which are critical determinants of structural compatibility and functional binding specificity in biological systems (12, 13). Furthermore, predictive models based on single classifiers often struggle to learn the complex and nonlinear interaction patterns inherent in host–virus systems, where viral proteins frequently target multiple host pathways simultaneously (14). These biological characteristics increase the functional complexity of PPI prediction tasks and limit the generalizability of conventional modeling strategies.

To overcome these challenges, we propose a structured ensemble learning framework for sequence-based prediction of human–virus PPIs. In this framework, protein sequences are transformed into continuous vector representations using Word2Vec-based embeddings, which capture contextual relationships between sequence fragments in a biologically meaningful manner. Context-dependent residue interactions and short linear motifs are known to play essential roles in host–pathogen binding specificity (15, 16). Embedding-based representations provide a scalable mechanism for encoding such sequence context without relying solely on handcrafted descriptors. These learned embeddings serve as input to multiple ensemble learning strategies, including boosting-based and bagging-based algorithms. By evaluating these ensemble paradigms within a unified representation framework, we systematically assess their ability to capture interaction-relevant sequence features.

Unlike approaches that rely on a single predictive model, our framework integrates complementary ensemble strategies to enhance predictive stability and accuracy. Boosting methods iteratively refine predictions by focusing on misclassified samples, enabling improved detection of subtle interaction

patterns, while bagging methods reduce variance through independent model aggregation. Viral proteins are known to exploit transient or weak interactions to hijack host cellular machinery (17), suggesting that adaptive modeling strategies capable of refining decision boundaries may be advantageous. Through comprehensive comparative evaluation, we demonstrate that ensemble learning particularly boosting-based methods provides improved robustness and generalization for embedding-driven PPI prediction.

Taken together, this work establishes an embedding-based ensemble learning framework that combines automated sequence representation with systematic comparative modeling. By integrating biologically meaningful embeddings with complementary ensemble strategies, the proposed approach provides a scalable and effective computational strategy for decoding complex human–virus interaction networks and supporting future antiviral research.

Advantages of the proposed ensemble framework

The primary contribution of this study lies in the development of an embedding-driven ensemble learning framework specifically designed to address the challenges associated with sequence-based prediction of human–virus PPIs. Unlike conventional approaches that apply a single classifier or focus on optimizing one specific ensemble algorithm, this work introduces a structured framework that integrates learned sequence embeddings with a systematic evaluation of multiple ensemble learning paradigms under identical experimental conditions. This design enables a direct investigation of how different ensemble strategies interact with biologically meaningful sequence representations, rather than confounding model performance with variations in feature engineering or data preprocessing. A key innovation of the proposed framework is the use of Word2Vec-based learned embeddings as the foundational representation for ensemble learning (18). These embeddings capture contextual relationships between sequence fragments, allowing proteins with similar functional or interaction characteristics to be represented in a shared feature space. However, such embedding-derived representations contain complex and distributed patterns that are not easily captured by conventional single-model classifiers (19). By leveraging ensemble learning, the proposed framework enables the extraction of these distributed interaction signals through multiple complementary decision structures. In particular, boosting-based models iteratively refine their predictions by focusing on previously misclassified samples (20,21), allowing the model to progressively learn difficult interaction patterns that may be overlooked during single-pass training.

Another distinctive strength of this work lies in its controlled comparative architecture. Rather than assuming the superiority of a specific ensemble algorithm, the framework evaluates multiple boosting and bagging methods within the same embedding-driven feature space. This allows for a clear and unbiased assessment of how each ensemble paradigm responds to biologically relevant sequence embeddings. The results demonstrate that boosting-based algorithms, particularly XGBoost, exhibit superior capability in capturing nonlinear and high-dimensional interaction signatures present in protein sequence embeddings (22). This advantage arises from their ability to adaptively refine decision boundaries and focus learning capacity on informative but difficult-to-classify interaction pairs.

Importantly, the proposed framework does not rely on handcrafted features or manually defined sequence descriptors. Instead, it provides a fully data-driven approach that integrates automated representation learning with ensemble-based predictive modelling (23). This combination improves predictive robustness, enhances generalization across diverse protein families, and reduces the risk of overfitting. From a methodological perspective, the novelty of this work lies not only in the use of ensemble learning, but in the systematic integration of embedding-based representation learning with controlled ensemble evaluation. This enables the framework to capture biologically meaningful interaction patterns more effectively while providing empirical insight into the relative strengths of different ensemble learning strategies for protein interaction prediction.

Comparative and competitive landscape analysis

The prediction of PPIs has been addressed using a wide range of computational approaches, including traditional machine learning classifiers, deep learning architectures, and ensemble-based methods. Each of these approaches offers specific advantages but also presents limitations when applied to high-dimensional sequence-derived representations and complex biological interaction patterns (24).

Traditional machine learning methods, such as support vector machines, k-nearest neighbors, and

standalone decision trees, rely on learning a single decision function from the available feature space. While these methods have demonstrated reasonable performance when combined with handcrafted features, they often struggle to capture the complex, nonlinear relationships inherent in protein interaction data. This limitation becomes particularly evident when using embedding-derived features, where interaction-relevant information is distributed across multiple dimensions rather than concentrated in a small number of dominant features. As a result, single-model classifiers may fail to fully exploit the representational richness of learned embeddings.

Deep learning models, including convolutional and recurrent neural networks, offer the ability to automatically learn hierarchical representations from sequence data and have shown promising results in interaction prediction tasks (25) However, their performance is strongly dependent on the availability of large training datasets and substantial computational resources. In many host–virus interaction scenarios, the available data are limited and highly imbalanced, which can restrict the effectiveness and stability of deep learning models (26). Additionally, the internal representations learned by deep neural networks are often difficult to interpret, limiting their transparency and practical applicability in biological research settings.

Ensemble learning methods provide a robust alternative by combining multiple predictive models to improve generalization and reduce prediction errors. Bagging-based approaches, such as random forest and extra trees, improve model stability by reducing variance through independent training on resampled datasets (27). However, because these models operate independently, their ability to iteratively refine decision boundaries (28) and capture subtle interaction patterns is inherently limited. Boosting-based methods, in contrast, improve predictive accuracy through sequential learning, where each model focuses on correcting the errors of its predecessors. This iterative refinement process enables boosting algorithms to identify weak but biologically meaningful interaction signals that may not be effectively captured by bagging methods or single classifiers.

The framework proposed in this study advances beyond existing approaches by integrating embedding-based sequence representation with a structured comparative evaluation of both boosting and bagging ensemble strategies. By evaluating these ensemble paradigms within a unified feature representation framework, this study provides direct empirical evidence of their relative effectiveness in learning interaction-relevant features from protein sequence embeddings. The observed superior performance of boosting algorithms, particularly XGBoost, demonstrates their enhanced ability to capture complex interaction patterns and improve prediction reliability.

Compared to traditional machine learning approaches, the proposed framework offers improved robustness and predictive accuracy by leveraging multiple complementary decision structures. Compared to deep learning methods, it provides competitive predictive performance while maintaining computational efficiency, interpretability, and practical applicability. Most importantly, by systematically evaluating multiple ensemble learning paradigms within a controlled and biologically meaningful representation framework, this study provides new methodological insight into the most effective strategies for embedding-driven protein interaction prediction. This combination of automated sequence representation, structured ensemble learning, and empirical comparative analysis establishes the proposed framework as a robust and effective approach for predicting human–virus PPIs, offering both practical predictive capability and methodological advancement over existing computational approaches.

Related work

Sequence-based approaches have become widely adopted for PPI prediction because they provide scalable and cost-effective alternatives to experimental discovery methods. These methods transform raw amino acid sequences into numerical descriptors that capture evolutionary, structural, and physicochemical information. For instance, the Substitution Matrix Representation (SMR) encodes evolutionary divergence through substitution rates in an N × 20 matrix (29,30) At the same time, the Multi-scale Local Descriptor (MLD) divides sequences into overlapping fragments to capture localized sequence patterns (14,15). Autocovariance descriptors, derived from Position-Specific Scoring Matrices (PSSMs), model positional dependencies across amino acids (31), whereas the Conjoint Triad (CT) representation reduces complexity by grouping amino acids into triplets, balancing interpretability and computational efficiency (32). These feature extraction techniques form the foundation for machine learning and deep learning models designed to capture the complexity of protein interactions.

Building upon these representations, a wide range of computational models has been investigated for PPI prediction. Traditional machine learning approaches such as Support Vector Machines (SVMs) and Random Forests (RFs) have shown strong predictive performance; however, they rely heavily on handcrafted features, which can limit scalability and generalization across diverse datasets. More recently, deep learning network models have enabled automated representation learning by capturing complex nonlinear dependencies within protein sequences. Representative examples include stacked autoencoder-based deep neural networks (DNNs), convolutional neural network–feature selective rotation forest (CNN-FSRF) models and Locality Preserving Projections (LPP) combined with Rotation Forest (RoF), as well as specialized frameworks such as Deep Neural Network with Local Conjoint Triad Descriptor (DNN-LCTD), Deep Neural Network for PPI prediction (DNN-PPI), Y-type Bidirectional Recurrent Neural Networks (Bi-RNNs), and Siamese-like Convolutional Neural Networks (CNNs) (33-45).

In parallel, ensemble learning method, including Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GB), Random Forest (RF), and Extremely Randomized Trees (Extra Trees) have gained considerable attention due to their ability to integrate multiple learners, reduce overfitting, and enhance predictive stability (44-50). These ensemble strategies provide an effective balance between accuracy, variance control, and computational efficiency, making them well-suited for modeling the nonlinear and heterogeneous characteristics of human–virus PPIs. Together, these methodological developments reflect a clear progression from handcrafted feature-based classifiers toward automated and ensemble-driven frameworks that improve both predictive robustness and biological relevance.

Materials and Methods

Proposed methodology

In this study, we propose a robust methodology to determine the most effective ensemble learning models for predicting PPIs by utilizing learned embeddings from both positive (interacting) and negative (non-interacting) samples.

Our approach involves a detailed evaluation and comparison of various ensemble techniques, focusing on boosting and bagging algorithms. The methodology is illustrated in Figure 1., which outlines the steps involved in predicting PPIs using embeddings derived from protein sequence data. The positive and negative datasets used in this study were retrieved from the Long Short-Term Memory for Protein-Human-Virus (LSTM-PHV) online web server, distinguishing between interacting and non-interacting proteins. We implemented an automated feature extraction method to facilitate machine learning model training, employing Word2Vec to generate learned embeddings that map protein sequences into a continuous vector space. This representation positions similar sequences closer together, improving the data quality. After generating the learned embeddings, the data was partitioned into training and testing sets. The training set was then used to train models using boosting and bagging ensemble techniques. This process involved applying various algorithms within these techniques to develop trained models for subsequent evaluation. Bagging, which trains multiple models independently and aggregates their predictions through averaging or voting, was executed using Extra Trees, Random Forest, and Decision Tree algorithms. In contrast, boosting sequential trains a series of weak learners and combines their predictions through weighted averaging to enhance performance, utilizing algorithms like XGBoost, LightGBM, and Gradient Boosting. Our evaluation of these ensemble models using the test set revealed that boosting algorithms, particularly XGBoost, exhibit superior effectiveness in predicting PPIs compared to bagging methods, highlighting their potential for achieving high accuracy and generalization in PPI prediction.

Protein sequence data

The present study utilizes a publicly available benchmark human–virus PPI dataset provided by the LSTM-PHV web server (51), originally derived from the Host–Pathogen Interaction Database 3.0 (HPIDB 3.0). The dataset was fully curated and preprocessed by the authors of LSTM-PHV using strict biological criteria to ensure reliability. Specifically, positive interactions were selected based on a confidence score threshold (MI score ≥ 0.3), redundancy removal using Cluster Database at High Identity with Tolerance (CD-HIT) at a 95% sequence identity threshold, and filtering to include only proteins composed of standard amino acids with lengths between 30 and 1000 residues. Negative samples were also generated using a dissimilarity-based sampling method, in which human–virus pairs sequence-similar to known interactors were excluded using the Needleman–Wunsch alignment algorithm with the BLOSUM30 substitution matrix. This ensured that the negative set consisted of curated non-interacting protein pairs rather than arbitrarily random pairs, thereby enhancing biological reliability. The final dataset

included 22,383 positive and 2,23,821 negative interactions.

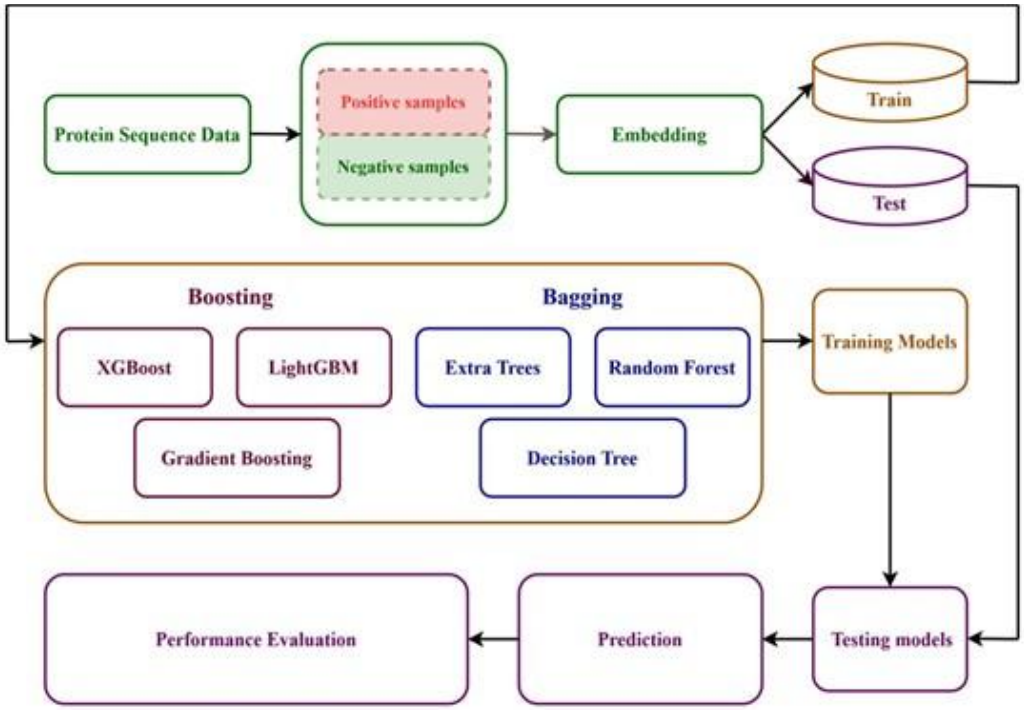


Figure 1. The architecture of the proposed methodology adopted for predicting PPIs using learned embeddings.

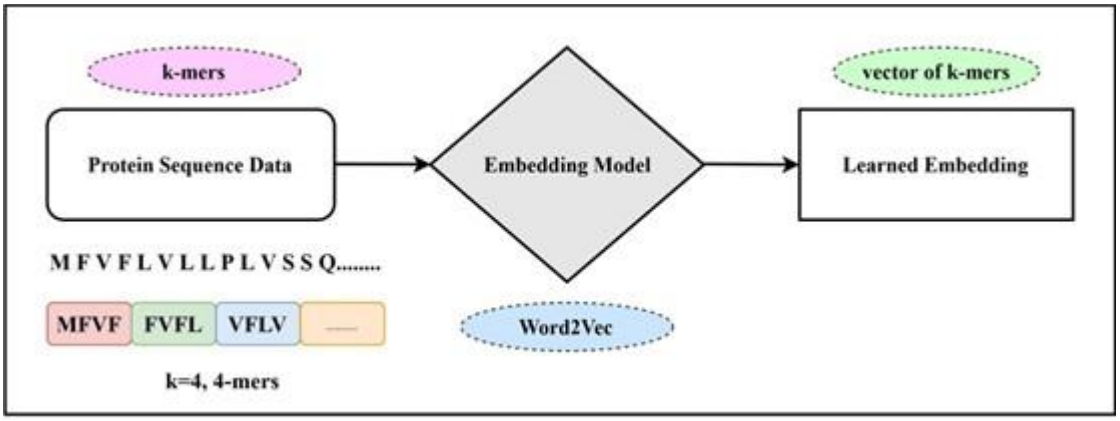


Figure 2. Creation of learned embeddings using Word2Vec embedding.

For our work, we adopted this standardized benchmark dataset to ensure comparability with previous studies. From the negative set, we have excluded very short or non-informative sequences to retain high-quality samples suitable for robust training. To address the inherent class imbalance, we then performed a random downsampling of the remaining negatives to obtain 22,383 samples, thereby maintaining a 1:1 ratio with the positive class. Importantly, since all negatives were already curated as non-interacting by LSTM-PHV, the random selection step was used purely for balancing and does not affect the biological validity or accuracy of the model. The downsampling was performed with a fixed random seed to guarantee reproducibility.

Embedding

Our proposed method converts protein sequence data into numerical representations by utilizing the Continuous Bag-of-Words (CBOW) model (52) from the Word2Vec framework to generate learned embeddings for k-mers. We begin by extracting fixed-length segments, known as k-mers, from the raw protein sequences. For example, we generate k-mers of length four, such as MFVF, FVFL, VFLV, etc., from these sequences depicted in Figure 2. The extracted k-mers are fed into the CBOW model from the Word2Vec framework. This model predicts each k-mer based on the context of its surrounding k-mers within the sequence. By leveraging the information from neighboring k-mers, the CBOW

model accurately represents each k-mer, capturing the contextual relationships and patterns in the protein sequences. Through iterative training, the model fine-tunes the numerical vectors for each k-mer to reduce prediction errors. This iterative adjustment helps the model learn the underlying semantic and syntactic relationships between k-mers, resulting in dense vectors that capture these connections. The resulting learned embeddings translate complex amino acid sequences into a structured numerical form, preserving the essential patterns and relationships within the data and retaining critical sequence information.

Training and testing models

The major goal of our research is to find the best ensemble learning model for predicting PPIs using learned embeddings. Ensemble learning enhances prediction accuracy by combining multiple models, which helps capture a more comprehensive array of patterns and correct errors from individual models. We explored two widely used ensemble learning techniques bagging and boosting. In boosting, we start by dividing the dataset into subsets and training an initial weak learner on one of these subsets. After evaluating the performance of the learner, we adjust the weights of misclassified data points so that subsequent models focus more on correcting these errors. This iterative process continues, with each new model improving on the errors of the previous ones, and predictions are aggregated to form a robust final model. In contrast, bagging involves creating multiple bootstrap samples from the original dataset. Each is used to train a separate weak learner independently. These models then make predictions combined through methods like majority voting or averaging. This approach helps to reduce variance and improve model stability. To evaluate the effectiveness of these models, we conducted a rigorous testing process on a separate dataset with positive and negative samples. This step is essential to understand how well the models generalize to new, unseen data and their accuracy in predicting protein interactions. Our results, derived from a comprehensive evaluation, indicate that boosting algorithms, especially XGBoost, outperform bagging models, demonstrating the effectiveness of boosting techniques in utilizing learned embeddings for accurate PPI prediction.

Results and Discussion

The proposed methodology for predicting PPIs employed a structured approach leveraging learned embeddings from protein sequence data. Positive and negative datasets were retrieved from the LSTM-PHV web server, where positive samples represented interacting proteins and negative samples represented non-interacting proteins. Protein sequences are represented as strings of amino acids. To generate word embeddings from this Word2Vec, a widely used tool in natural language processing, was applied to extract features from protein sequence data. When applied, Word2Vec first requires the sequences to be preprocessed, typically by tokenizing them into individual amino acids. The model is then trained using a Continuous Bag of Words (CBOW), which predicts the current amino acid based on its context. After training, Word2Vec generates vector embeddings for each amino acid, reflecting their semantic relationships and positioning similar amino acids closer in vector space. These embeddings are then used to represent protein sequences as sequences of vectors. Fixed feature vectors are generated for each protein called learned embeddings by averaging these learned embeddings. Now consisting of learned embeddings, the dataset is divided into training and testing sets and subjected to boosting and bagging ensemble techniques. The study compared boosting algorithms (XGBoost, LightGBM, and Gradient Boosting) and bagging algorithms (Extra Trees, Random Forest, and Decision Tree) for PPI prediction. The analysis revealed that boosting methods outperformed bagging techniques across all evaluation metrics, including accuracy, precision, recall, and F1 score. XGBoost emerged as the top-performing algorithm, achieving the highest accuracy and balanced performance in precision, recall, and F1 scores, effectively minimizing false positives and negatives, as shown in Table 1.

These findings highlight the advantages of boosting algorithms, particularly XGBoost, in PPI prediction. Boosting's iterative training, where each model corrects the errors of the previous one, allows it to adapt more effectively to the intricate patterns within PPI data. In contrast, bagging methods, though robust, showed less effectiveness in handling the complexity of non-linear biological interactions. This study underscores the importance of algorithm selection in biological data analysis, suggesting that boosting techniques, particularly XGBoost, are well-suited for tasks involving complex datasets such as PPIs. The results reinforce the superiority of boosting approaches in PPI prediction due to their error correction capabilities, offering valuable insights for bioinformatics research

Figure 3. presents the confusion matrices for XGBoost, LightGBM, and Gradient Boosting, illustrating their classification performance in predicting PPIs. Among the three, XGBoost exhibits the highest accuracy, with a greater concentration of correct classifications and fewer misclassifications, particularly in reducing errors where class 1 is misclassified as class 0. LightGBM also performs

Table 1. Comparative analysis of evaluation metrics for boosting and bagging algorithms.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Boosting Algorithm				
XGBoost	89.05	89.85	89.83	89.05
LightGBM	84.54	84.77	84.52	84.54
Gradient Boosting	80.83	80.85	80.83	80.83
Bagging Algorithm				
Extra Trees	83.60	83.61	83.60	83.60
Random Forest	82.36	82.38	82.36	82.36
Decision Tree	75.41	75.54	75.42	75.41

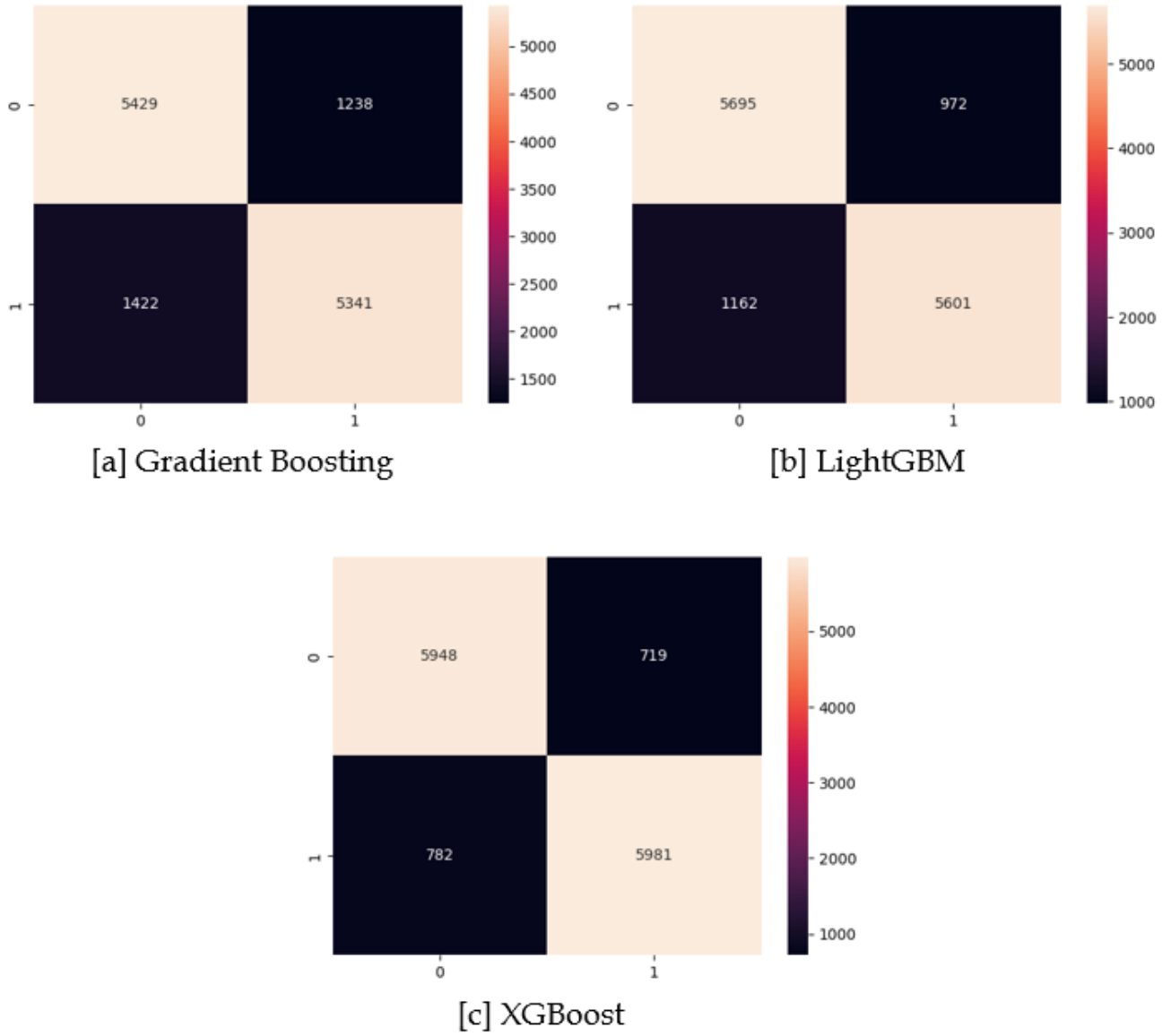


Figure 3. Confusion matrix for the boosting algorithm.

well, though with a slight increase in misclassifications compared to XGBoost, reflecting a marginal reduction in accuracy. Gradient Boosting shows the highest misclassifications, indicating relatively lower performance than the other two models. Figure 4. illustrates the confusion matrices for Extra Trees, Random Forest, and Decision Tree bagging algorithms, highlighting their classification performance in predicting PPIs. Extra Trees demonstrates strong performance with high true positive (tp) and true negative (tn) rates, reflecting its effectiveness in accurately classifying both classes. Its low false positive (fp) and false negative (fn) rates further suggest minimal misclassifications,

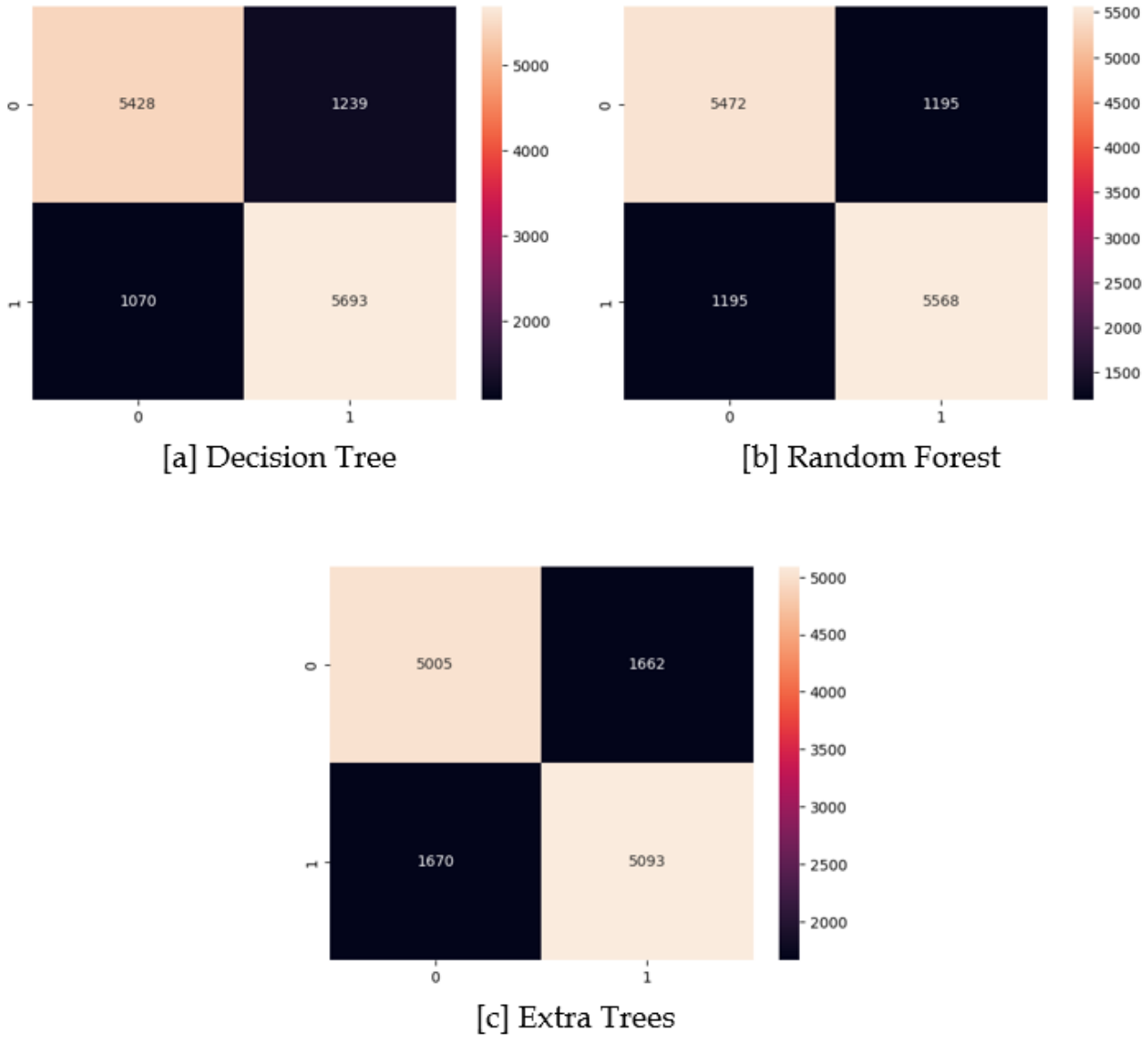


Figure 4. Confusion matrix for the bagging algorithm.

resulting in a well-balanced sensitivity and specificity. Similarly, Random Forest performs comparably to Extra Trees, showing high tp and tn values with slight differences in accuracy. In contrast, the Decision Tree exhibits lower tp and tn rates alongside higher fp and fn rates, indicating greater misclassifications and challenges in capturing complex patterns. As a result, Extra Trees and Random Forest exhibit superior classification accuracy and better error reduction than the Decision Tree.

The ROC curves for bagging and boosting algorithms in Figure 5. further corroborate the superior performance of XGBoost, showing the highest Area Under the Curve (AUC) scores, demonstrating better discrimination between positive and negative PPIs. LightGBM and Gradient Boosting, while exhibiting similar performance with overlapping ROC curves, achieve slightly lower AUC values than XGBoost. Among the bagging methods Extra Trees and Random Forest display similar ROC curves, closely approaching the top-left corner, indicating superior classification accuracy and a favorable trade-off between true positive rate (tpr) and false positive rate (fpr). In contrast, the Decision Tree displays a lower ROC curve and reduced AUC, indicating weaker discriminatory ability.

Comparing results of boosting and bagging algorithms

The comparison of Boosting and Bagging algorithms was conducted using several key performance indicators, including ROC curves, training time, AUROC scores, sensitivity, specificity, and overall classification metrics. Figures 6. and 7. comprehensively evaluate the performance of six classification algorithms Decision Tree, Random Forest, Extra Trees, Gradient Boosting, LightGBM,

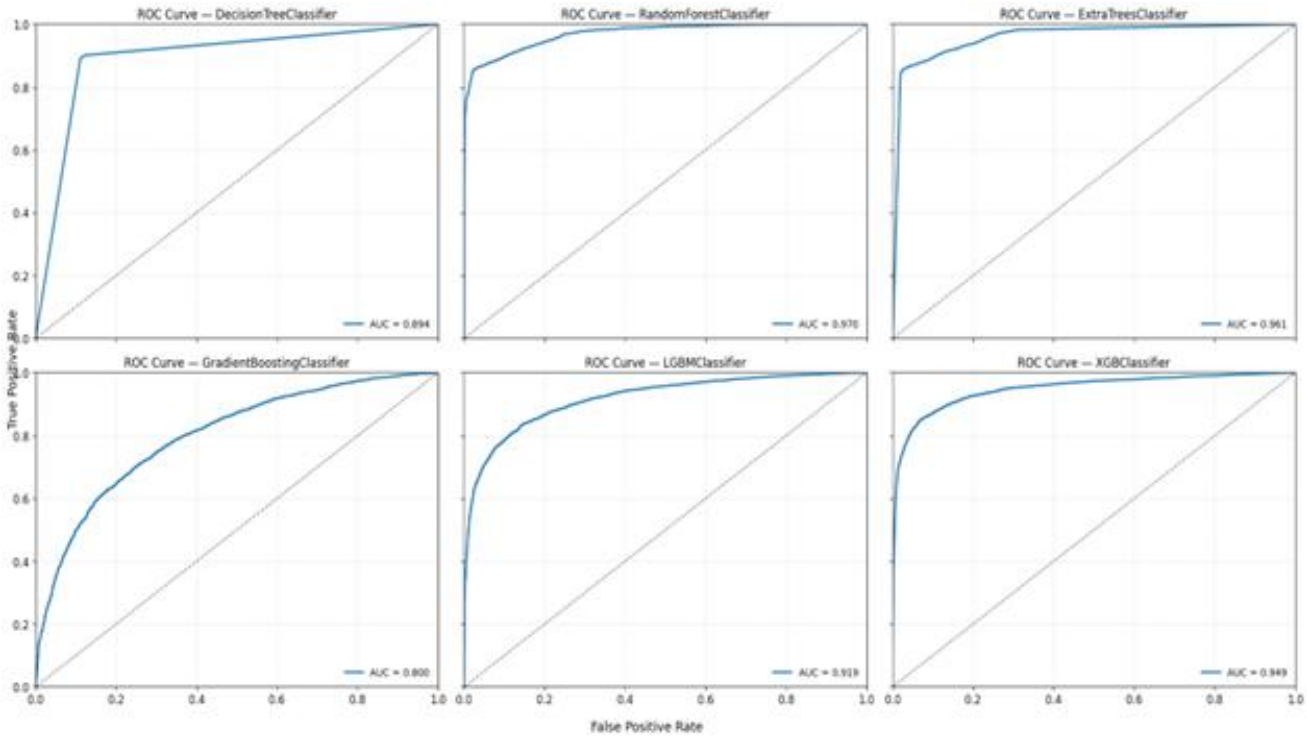


Figure 5. ROC curve for the bagging and boosting algorithms.

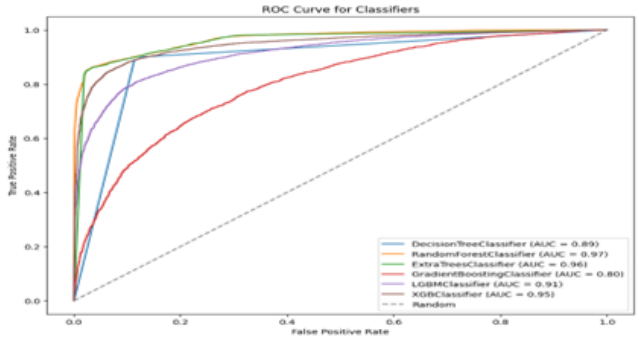


Figure 6. ROC curves illustrating the trade-off between true positive rate and false positive rate for each classifier.

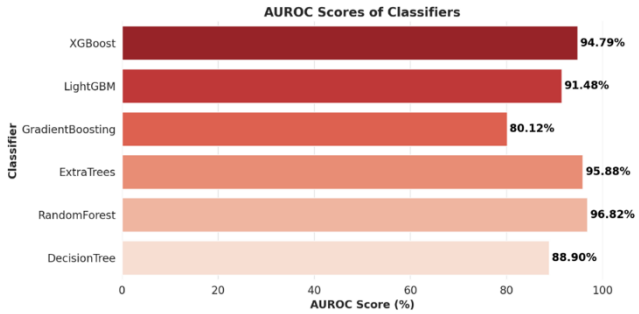


Figure 7. Classifier performance comparison based on AUROC values.

and XGBoost through ROC curves and AUROC scores. Figure 7. illustrates that XGBoost and Random Forest achieve the highest ROC curves, indicating superior classification accuracy and an effective balance between true positive rate (tpr) and false positive rate (fpr). Extra Trees also demonstrates strong performance, with a ROC curve that closely aligns with XGBoost and Random Forest. In contrast, Gradient Boosting, LightGBM, and Decision Tree exhibit lower ROC curves, suggesting comparatively reduced effectiveness in distinguishing between positive and negative classes. Figure 8. complements this analysis by presenting the area under the ROC curve (AUROC) scores, which quantify overall classifier performance. XGBoost leads with the highest AUROC score, reaffirming its exceptional classification capabilities. Random Forest and Extra Trees also achieve high AUROC scores, reflecting their strong performance. Conversely, Gradient Boosting, LightGBM, and Decision Tree score lower on the AUROC metric, indicating that these models are less effective for this task.

Figure 8. demonstrates that Gradient Boosting is the most computationally intensive, consuming approximately 70% of the total training time. In contrast, Decision Tree, Random Forest, Extra Trees, LightGBM, and XGBoost exhibit significantly lower and comparable training times, ranging from 4% to 10% of the total. Figure 9. integrates training time with AUROC scores. This metric measures the area under the receiver operating characteristic curve and is a key indicator of the ability of the model to distinguish between classes. This figure reveals that while Gradient Boosting achieves the highest AUROC score, it is also the most resource-demanding. XGBoost, however, offers a favorable balance by delivering high performance with a relatively lower training time. Conversely, Decision Tree, Random Forest, Extra Trees, and LightGBM, while requiring less training time, show lower

AUROC scores compared to Gradient Boosting and XGBoost. These findings suggest that Gradient

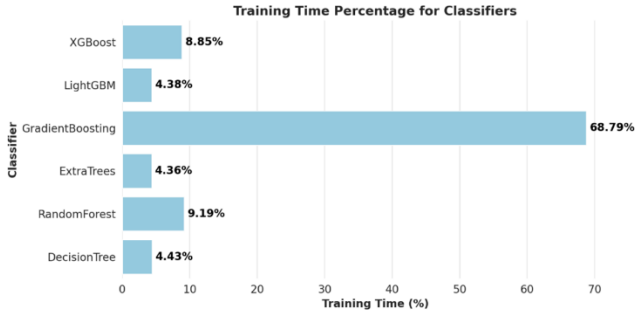


Figure 8. Bar graph illustrating the proportion of training time required by different classification algorithms.

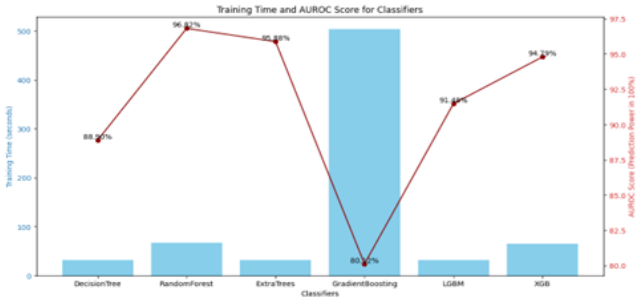


Figure 9. Visualization of classifier performance based on AUROC scores and training time.

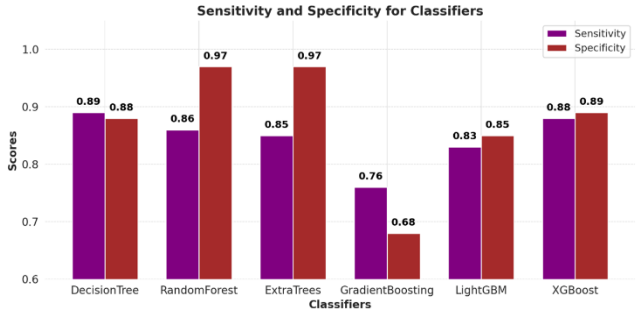


Figure 10. Evaluation of classifier performance based on sensitivity and specificity metrics.

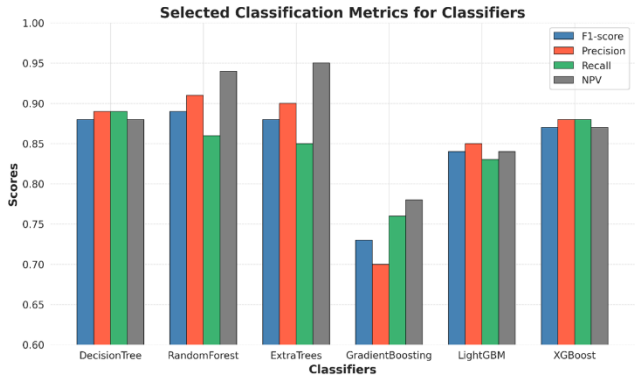


Figure 11. Bar chart illustrating the classification metrics of various machine learning models.

Boosting excels in classification accuracy but at a higher computational cost, whereas XGBoost presents a practical compromise between performance and efficiency. Figure 10. illustrates that XGBoost achieves the highest sensitivity and specificity, reflecting its superior capacity for correctly classifying positive and negative instances. Random Forest, Extra Trees, and Gradient Boosting also exhibit strong sensitivity and specificity, although their performance varies slightly compared to XGBoost. In contrast, Decision Tree and LightGBM exhibit lower sensitivity and specificity, indicating potential limitations in capturing certain patterns in the data.

Figure 11. overviews additional classification metrics, including F1-score, recall, precision, and accuracy. XGBoost excels across all these metrics, underscoring its superior classification task performance. Random Forest, Extra Trees, and Gradient Boosting also achieve high scores in most metrics, indicating robust classification capabilities. However, Decision Tree and LightGBM show lower scores in several areas, suggesting limitations in balancing precision and recall or in capturing specific patterns. Collectively, these results highlight XGBoost as the most effective classifier based on sensitivity, specificity, and other critical performance metrics.

The analysis highlights that boosting algorithms, particularly XGBoost, perform better in predicting PPIs than other classification methods. XGBoost achieves the highest ROC curves and AUROC scores and demonstrates exceptional classification accuracy and effective discrimination between positive and negative classes. Its ability to handle complex patterns and interactions within the data makes it a robust choice for this predictive task. Therefore, XGBoost is recommended as the optimal algorithm for enhancing prediction accuracy in this domain, providing a significant advantage over other classifiers.

Limitations and future directions

In our earlier work [53], we observed that traditional composition-based feature descriptors required extensive manual feature engineering and were limited in their ability to capture contextual relationships between amino acids. This methodological limitation motivated the transition toward an embedding-driven representation framework in the present study.

By adopting a Word2Vec-based embedding strategy, we introduced an automated and data-driven mechanism for representing protein sequences. The 4-mer embeddings capture local contextual relationships between sequence fragments and reduce dependence on predefined heuristic descriptors. In comparison to handcrafted

feature-based approaches and single-model classifiers, this embedding-driven ensemble framework provides improved scalability, better generalization across diverse protein families, and enhanced ability to extract distributed interaction-relevant patterns. Furthermore, by mapping influential embedding features back to their originating k-mers, we introduced an additional level of interpretability, allowing identification of biologically meaningful motifs associated with receptor binding and immune modulation. This represents a methodological advancement over earlier feature-engineering-based approaches.

Nevertheless, the proposed embedding strategy has inherent limitations. Representing each protein sequence through averaged embeddings of overlapping k-mers, while computationally efficient, inevitably results in the loss of positional and higher-order contextual information. The averaging process treats k-mers as position-independent units, potentially obscuring sequential dependencies and long-range interaction motifs that may play critical roles in protein–protein interactions. While ensemble learning enhances pattern extraction from the embedding space, it cannot fully compensate for the representational constraints imposed by static averaging.

When positioned within the broader methodological landscape, this study advances beyond handcrafted feature models and single-classifier frameworks by integrating automated embeddings with systematic ensemble evaluation. However, it does not yet incorporate sequence-aware deep architectures capable of explicitly modeling long-range dependencies. Modern sequence modeling approaches, including recurrent neural networks (RNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), convolutional neural networks (CNNs), and transformer-based architectures with self-attention mechanisms, are inherently designed to preserve positional and contextual information across entire sequences. These models may further enhance predictive performance by capturing interaction signals that static embedding representations cannot fully encode.

Future work should therefore explore the integration of context-preserving neural architectures with ensemble learning strategies. Hybrid frameworks combining transformer-based embeddings with boosting-based classifiers, or sequence-aware deep models evaluated within controlled ensemble paradigms, may provide a balanced trade-off between representational richness and predictive robustness. Such integration would extend the current framework by addressing positional information loss while maintaining interpretability and computational practicality.

Finally, it is important to emphasize that the present study is entirely computational and functions as a hypothesis-generating analysis. The predicted interactions arise from machine learning models trained on sequence-derived representations and have not been experimentally validated. While the computational results provide valuable prioritization of potential host–virus interaction pairs, experimental validation through biochemical assays, structural analysis, or cellular studies is necessary to confirm biological relevance. Future collaboration with experimental laboratories will be essential to translate these predictions into validated mechanistic insights.

Conclusion

This study presents a structured ensemble learning framework for the prediction of human–virus protein–protein interactions using embedding-based sequence representations. By integrating Word2Vec-derived k-mer embeddings with systematic evaluation of both boosting and bagging ensemble paradigms, we established a controlled comparative framework for assessing model behavior on biologically meaningful sequence features. The results consistently demonstrate that boosting-based methods, particularly XGBoost, outperform bagging strategies across multiple evaluation metrics, indicating their superior capacity to extract distributed and nonlinear interaction signals embedded within protein sequence representations. A key contribution of this work lies in the combination of automated feature learning and structured ensemble comparison. Unlike traditional approaches dependent on handcrafted descriptors or single classifiers, the proposed framework leverages contextual sequence embeddings while maintaining computational efficiency and interpretability. The findings suggest that gradient-based ensemble strategies are particularly well suited for modeling embedding-derived biological data, where interaction-relevant signals are subtle and high-dimensional. At the same time, the framework remains scalable and accessible, providing a practical alternative to resource-intensive deep learning architectures. Importantly, this study advances methodological clarity by systematically evaluating how different ensemble learning paradigms interact with learned sequence embeddings under identical experimental conditions. Beyond improving predictive accuracy, the work offers guidance for selecting appropriate modeling strategies for host–virus interaction prediction. In the broader context

of infectious disease research, embedding-driven ensemble models can serve as effective computational tools for prioritizing candidate host–pathogen interactions, supporting hypothesis generation, and accelerating downstream experimental validation. Together, these contributions reinforce the role of structured, interpretable machine learning frameworks in advancing computational virology and systems biology.

Contribution of Authors

The study was conceived and designed by Sini S Raj. Sini S Raj also managed material preparation, data collection, and analysis. The first draft of the manuscript was written by Sini S Raj and critically reviewed by Vinod Chandra S S. Both authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to thank the researchers and the staff members of the Machine Intelligence Research Lab at the Department of Computer Science, University of Kerala, for providing all the facilities and support for carrying out research of this scale.

Conflict of Interest

All authors declare that they have no conflicts of interest.

Funding

No funding was received for this study.

Data Availability

The authors declare that all the data used in the current study are available upon request from the corresponding author.

References

1. Xu, X.; Chen, P.; Wang, J.; Feng, J.; Zhou, H.; Li, X.; Zhong, W.; Hao, P. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Science China Life Sciences 2020, 63, 457–460.
2. Wu, J.; Yuan, X.; Wang, B.; Gu, R.; Li, W.; Xiang, X.; Tang, L.; Sun, H. Severe acute respiratory syndrome coronavirus 2: from gene structure to pathogenic mechanisms and potential therapy. Frontiers in Microbiology 2020, 11, 1576.
3. Hu, B.; Guo, H.; Zhou, P.; Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology 2021, 19 (3), 141–154.
4. Trottein, F.; Sokol, H. Potential causes and consequences of gastrointestinal disorders during a SARS-CoV-2 infection. Cell Reports 2020, 32 (3).
5. Decaro, N.; Lorusso, A. Novel human coronavirus (SARS-CoV-2): A lesson from animal coronaviruses. Veterinary Microbiology 2020, 244, 108693.
6. Esposito, M. M.; Turku, S.; Lehrfield, L.; Shoman, A. The Impact of Human Activities on Zoonotic Infection Transmissions. Animals 2023, 13 (10), 1646.
7. Dyer, M. D.; Murali, T. M.; Sobral, B. W. The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathogens 2008, 4 (2), e32.
8. Lian, X.; Yang, X.; Yang, S.; Zhang, Z. Current status and future perspectives of computational studies on human–virus protein–protein interactions. Briefings in Bioinformatics 2021, 22 (5), bbab029.
9. De Las Rivas, J.; Fontanillo, C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Computational Biology 2010, 6 (6), e1000807.
10. Yi, B.; Deng, Q.; Guo, C.; Li, X.; Wu, Q.; Zha, R.; Wang, X.; Lu, J. Evaluating the zoonotic potential of RNA viromes of rodents provides new insight into rodent-borne zoonotic pathogens in Guangdong, China. One Health 2023, 17, 100631.
11. Duffy, S. Why are RNA virus mutation rates so damn high? PLoS Biology 2018, 16 (8), e3000003.
12. Luck, K.; Kim, D.-K.; Lambourne, L.; Spirohn, K.; Begg, B. E.; Bian, W.; et al. A reference map of the human binary protein interactome. Nature 2020, 580 (7803), 402–408.
13. Gordon, D. E.; Jang, G. M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K. M.; et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020, 583 (7816), 459–468.
14. Via, A.; Uyar, B.; Brun, C.; Zanzoni, A. How pathogens use linear motifs to perturb host cell networks. Trends in Biochemical Sciences 2015, 40 (1), 36–48.
15. Weatheritt, R. J.; Gibson, T. J. Linear motifs: lost in (pre)translation. Trends in Biochemical Sciences 2012, 37 (8), 333–341.
16. Calderwood, M. A.; Venkatesan, K.; Xing, L.; Chase, M. R.; Vazquez, A.; Holthaus, A. M.; et al. Epstein–Barr virus and virus human protein interaction maps. Cell 2007, 130 (5),

889–899.
17. Asgari, E.; Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 2015, 10 (11), e0141287.
18. Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015, 16 (6), 321–332.
19. Schapire, R. E. The strength of weak learnability. Machine Learning 1990, 5, 197–227.
20. Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 1997, 55 (1), 119–139.
21. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, August 13–17, 2016; pp 785–794.
22. Vidal, M.; Cusick, M. E.; Barabási, A.-L. Interactome networks and human disease. Nature Reviews Genetics 2011, 12 (9), 615–628.
23. Hashemifar, S.; Neyshabur, B.; Khan, A. A.; Xu, J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics 2018, 34 (17), i802–i810.
24. Chen, H.; Li, F.; Wang, L.; Jin, Y.; Kurgan, L. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. Briefings in Bioinformatics 2021, 22 (3), bbaa068.
25. Breiman, L. Random forests. Machine Learning 2001, 45, 5–32.
26. Huang, Y.-A.; You, Z.-H.; Gao, X.; Wong, L.; Wang, L. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. BioMed Research International 2015, 2015 (1), 902198.
27. Huang, Y.-A.; You, Z.-H.; Li, X.; Chen, X.; Hu, P.; Li, S.; Luo, X. Construction of reliable protein–protein interaction networks using weighted sparse representation-based classifier with pseudo substitution matrix representation features. Neurocomputing 2016, 218, 131–138.
28. Zamil, K. S.; Rahman, J. Prediction of protein-protein interaction from amino acid sequence using ensemble classifier. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, February 8, 2018; IEEE: 2018; pp 1–4.
29. Brown, C. A.; Hansen, H. N.; Jiang, X. J.; Blateyron, F.; Berglund, J.; Senin, N.; Bartkowiak, T.; Dixon, B.; Le Goïc, G.; Quinsat, Y.; Stemp, W. J. Multiscale analyses and characterizations of surface topographies. CIRP Annals 2018, 67 (2), 839–862.
30. Yang, X.; Yang, S.; Li, Q.; Wuchty, S.; Zhang, Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and Structural Biotechnology Journal 2020, 18, 153–161.
31. Chen, H.; Li, F.; Wang, L.; Jin, Y.; Chi, C.-H.; Kurgan, L.; Song, J.; Shen, J. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. Briefings in Bioinformatics 2021, 22 (3), bbaa068.
32. Alashwal, H.; Deris, S.; Othman, R. M. One-class support vector machines for protein-protein interactions prediction. International Journal of Biological and Medical Sciences 2006, 1 (2), 120–127.
33. Chen, X.-W.; Liu, M. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 2005, 21 (24), 4394–4400.
34. Jensen, L. J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; Bork, P. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 2009, 37 (suppl_1), D412–D416.
35. Wang, L.; Wang, H. F.; Liu, S. R.; Yan, X.; Song, K. J. Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. Scientific Reports 2019, 9 (1), 9848.
36. Zhan, X.; You, Z.; Yu, C.; Pan, J.; Li, R. Predicting Protein-Protein Interactions from Protein Sequence Using Locality Preserving Projections and Rotation Forest. In Intelligent Computing Theories and Application: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part II; Springer International Publishing: 2020; pp 121–131.
37. Wang, J.; Zhang, L.; Jia, L.; Ren, Y.; Yu, G. Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. International Journal of Molecular Sciences 2017, 18 (11), 2373.
38. Gui, Y.; Wang, R.; Wei, Y.; Wang, X. DNN-PPI: a large-scale prediction of protein–protein interactions based on deep neural

networks. Journal of Biological Systems 2019, 27 (01), 1–8.
39. Yang, L.; Han, Y.; Zhang, H.; Li, W.; Dai, Y. Prediction of Protein-Protein Interactions with Local Weight-Sharing Mechanism in Deep Learning. BioMed Research International 2020, 2020 (1), 5072520.
40. Hashemifar, S.; Neyshabur, B.; Khan, A. A.; Xu, J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics 2018, 34 (17), i802–i810.
41. Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics 2020, 8 (5), 765.
42. Nobre, J.; Neves, R. F. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications 2019, 125, 181–194.
43. Taha, A. A.; Malebary, S. J. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access 2020, 8, 25579–25587.
44. Song, J.; Liu, G.; Jiang, J.; Zhang, P.; Liang, Y. Prediction of protein–ATP binding residues based on ensemble of deep convolutional neural networks and LightGBM algorithm. International Journal of Molecular Sciences 2021, 22 (2), 939.
45. Song, J.; Liu, G.; Jiang, J.; Zhang, P.; Liang, Y. Prediction of protein–ATP binding residues based on ensemble of deep convolutional neural networks and LightGBM algorithm. International Journal of Molecular Sciences 2021, 22 (2), 939.
46. Breiman, L. Bagging predictors. Machine Learning 1996, 24, 123–140.
47. Alelyani, S. Stable bagging feature selection on medical data. Journal of Big Data 2021, 8 (1), 11.
48. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. Machine Learning 2006, 63, 3–42.
49. Rodriguez, J. J.; Kuncheva, L. I.; Alonso, C. J. Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 2006, 28 (10), 1619–1630.
50. Hong, S.; Lynn, H. S. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology 2020, 20, 199.
51. Tsukiyama, S.; Hasan, M. M.; Fujii, S.; Kurata, H. LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. Briefings in Bioinformatics 2021, 22 (6), bbab228.
52. Hamid, M.-N.; Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics 2019, 35 (12), 2009–2016.
53. Sini Raj, S.; Vinod Chandra, S. S. Significance of sequence features in classification of protein–protein interactions using machine learning. The Protein Journal 2024, 43 (1), 72–83.