SEEJPH Volume XXVI, S2,2025, ISSN: 2197-5248; Posted:03-02-25

Synergistic Ensemble Methods for Predicting VEGF Sequences Associated with Dental Cystic Lesions and Ameloblastomas

Dr. Nausathkhan Ubayathulla [1], Dr. M.R. Muthusekar [2*], Dr. Prathiba Ramani [3], Dr. Pradeep Kumar Yadalam [4], Dr. Subasree S [5]

[1] Phd Scholar, Saveetha Dental College and Research Institute, Saveetha University, SIMATS, Chennai, Tamil Nadu, India Consultant Oral &Maxillofacial surgeon, EHS, Fujairah Specialized Dental Center and Hospital, Fujairah, UAE, Email id : nausathkhan@gmail.com

[2] Program Director, Saveetha Dental College and Research Institute, Saveetha University, SIMATS, Chennai, Tamil Nadu, India Email id : muthusekhar55@gmail.com

[3] HOD, Department of Oral and Maxillofacial Pathology, Saveetha Dental College and Research Institute, Saveetha University, SIMATS, Chennai, Tamil Nadu, India, Email id : pratibaramani@saveetha.com

[4] HOD, Department of Periodontology, Saveetha Dental College and Research Institute, Saveetha University, SIMATS, Chennai, Tamil Nadu, India, pradeepkumar.sdc@saveetha.com

[5] Assistant Professor, Department of Periodontology, Saveetha Dental College and Research Institute, Saveetha University, SIMATS, Chennai, Tamil Nadu, India, subasrees.sdc@saveetha.com

**Corresponding Author: Dr. M.R. Muthusekar MDS*

KEYWORDS

Dentigerous cyst, ameloblastoma, VEGF, ensemble learning.

ABSTRACT

Introduction: VEGF/VPF is a 45kd glycoprotein found on human chromosome 6p21.3 that binds to endothelial cell receptors. Five forms of the VEGF gene have been identified in mammals, promoting endothelial cell proliferation and migration and contributing to normal angiogenesis and pathological conditions like tumorigenesis. Targeting VEGF receptors is crucial for treating dentigerous cysts and tumors, requiring a sequence-based understanding for targeted therapies and surgical success. The study explores the role of VEGF in dental cystic lesions and Ameloblastomas, aiming to improve clinical outcomes by understanding tumor behavior. It uses ensemble methods for predicting VEGF sequences, integrating multi-omics data and machine learning to provide insights into molecular pathology and improve patient outcomes.

Methods: The human VEGF sequences, including P15692, A0A0A0MTB2, |P49765, O43915, A0A0A0MSI7, A0A0A0MRQ4, and Q7LAP4, were retrieved, checked for missing values, and analyzed for prediction using Protbert embeddings and stacked ensemble learning. The study used a dataset of 1,034 features from advanced measurements to enhance signal-to-noise ratio, reduce dimensionality, and retain critical information. The dataset was preprocessed using feature scaling and a hybrid feature selection method. The Standard Scaler standardized features, while the XGBoost algorithm refined selection. The final set of 459 optimized features was combined through stacking.

Results: The stacked ensemble model achieved 70% accuracy, relying heavily on gradient-boosting techniques. XGBoost optimizes memory usage and handles missing data. Random Forest may not effectively extract dataset complexities. Classifier weight analysis guides future strategies, focusing on boosting methods.

Conclusion: The study developed an ensemble model to predict VEGF-related outcomes in ameloblastoma and dentigerous cysts patients, but challenges like data imbalance and overfitting need to be addressed.

1. Introduction

VEGF/VPF is a 45 kd, heparin-binding glycoprotein found on human chromosome 6p21.3, specifically binding to endothelial cell receptors(1,2). Five VEGF forms in mammals have been identified through alternative splicing of the single VEGF gene: VEGF121, VEGF145, VEGF165, VEGF189, and VEGF206.(3–5)Three VEGF tyrosine kinase receptors, VEGFR-1/Flt, VEGFR-2/KDR/Flk-1, and VEGFR-3/Flt-4, have immunoglobulin-like regions. Angiogenesis is the formation of new blood vessels from existing vasculature, crucial for inflammation and tumor growth. The VEGF family regulates it, stimulating endothelial cell proliferation and migration, contributing to normal

angiogenesis and pathological conditions like tumorigenesis.(1).

Vascular development involves two stages: vasculogenesis, which forms the primary capillary network from mesoderm-derived precursors, and angiogenesis, which involves the formation of new capillary blood vessels from preexisting microvessels.(6). Vasculogenesis involves differentiation, proliferation, and coalescence, while angiogenesis involves remodeling the primary plexus to form new microcirculations. Angiopoietins, vascular endothelial cell-specific growth factors, play important roles in the later stages of angiogenesis. The VEGF-VEGFR complex promotes endothelial cell activity, and alterations in this balance affect proliferation and migration. Mucin 1 (MUC1) overexpression activates pro-angiogenic pathways, while decreased MUC1-C levels inhibit apoptosis and promote angiogenesis, affecting cancer metastasis.(1). Oncogenes, suppressor genes, growth factors, promoters, and hypoxia influence VEGF expression.(5,6).

One previous study explored the role of JunB and VEGF in angiogenesis in odontogenic keratocysts and dentigerous cysts. Results show higher JunB expression in OKCs (74.6%) and VEGF expression in DCs (52.88%)(7), indicating a significant correlation. Another previous study explored VEGF expression in ameloblastoma and dentigerous cysts, revealing varying levels in Ameloblastoma and a higher level in Dentigerous Cysts. This suggests VEGF's role in invasion of aggressive odontogenic lesions and suggests potential diagnostic and prognostic markers. These studies explore the presence and activity of VEGF in clinical samples, and targeting these VEGF receptors is crucial for treating dentigerous cysts and tumors.(2). Understanding this growth factor from a sequence point of view is necessary for targeted therapies and surgical success.

VEGF expression is upregulated in oral tumors like dental cystic lesions and ameloblastomas, causing altered angiogenesis. Understanding VEGF regulation is crucial for developing targeted therapies. Dentigerous cysts and ameloblastomas can cause severe morbidity if not managed properly, necessitating accurate diagnosis and treatment strategies. Research on predicting VEGF expression is essential for understanding these lesions' biological behavior. Bioinformatics primarily employs sequence-based computational models to tackle biological issues, but these models are limited to relying solely on sequence data(8–10). This work presents the computational model for accurate VEGF sequence prediction, addressing the drawbacks of current experimental methods like high costs and time requirements(10–13).

Predicting VEGF sequences helps researchers identify variations affecting expression and activity, elucidating angiogenesis and tumor progression pathways. One previous study showed that it developed a novel feature descriptor called MF-PSSM-DWT for generating primary sequence datasets. Deep learning techniques like GAN, GRU, ERCNN, and CNN were used for model training, achieving high accuracy rates.(14). Accurate VEGF therapeutic targeting prediction has revolutionized treatment techniques, enabling risk stratification and tailored interventions for patients in both health and disease. Ensemble methods(15) These computational techniques combine multiple predictive models for more accurate and robust predictions. This study uses synergistic ensemble methods to predict VEGF sequences, providing insights into the intricate role of VEGF in these conditions and potential therapeutic strategies. Ensemble methods have been successfully applied in bioinformatics, ecological modeling, and medicine. Applying ensemble methods to predict VEGF sequences could improve accuracy, particularly in dealing with noisy, variable, and sparse biological data.

The study takes a comprehensive approach to exploring the role of VEGF in dental cystic lesions and Ameloblastomas, with the ultimate goal of improving clinical outcomes by understanding tumor behavior. It investigates ensemble methods for predicting VEGF sequences, integrating multi-omics data and machine learning to provide a holistic view of the tumor microenvironment. The study suggests that ensemble methods for predicting VEGF sequences in dental cystic lesions and ameloblastomas could provide valuable insights into their molecular pathology. Identifying VEGF genetic variations could enhance diagnostic and treatment strategies, improve patient outcomes, and

demonstrate the potential of computational methods in molecular oncology in dentistry. Our study is a testament to the thoroughness of our approach, focusing on Synergistic Ensemble Methods for Predicting VEGF Sequences Associated with Dental Cystic Lesions and Ameloblastomas.

2. Methods

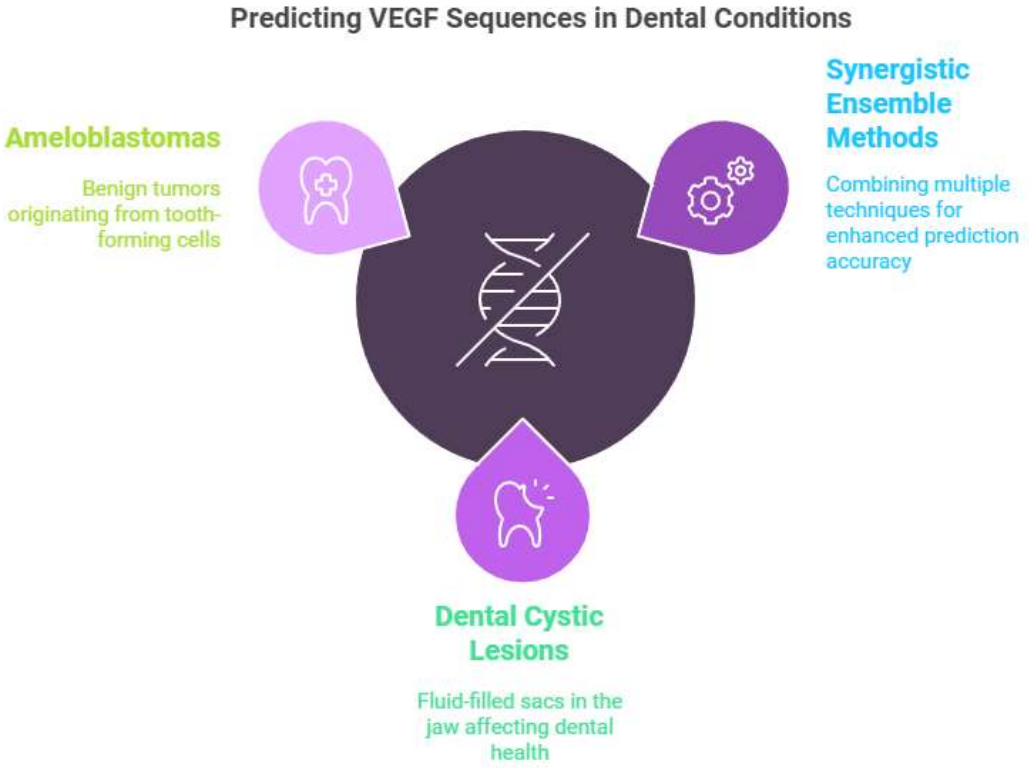

Figure 1 Shows the flowchart of the work model.

3. Data Retrieval

Using UNIPROT(16), FASTA sequences of VEGF of human origin- P15692|, A0A0A0MTB2, |P49765, O43915, A0A0A0MSI7, A0A0A0MRQ4, Q7LAP4, was retrieved and checked for missing values and duplicates and analyzed for prediction using Protbert embeddings with stacked ensemble learning. (fig-1)

Data Preprocessing and Feature Engineering

The dataset, comprising 1,034 features derived from advanced measurements, was subjected to a comprehensive preprocessing pipeline designed to enhance the signal-to-noise ratio, reduce dimensionality, and retain critical information. The study used feature scaling and a hybrid feature selection method to manage the high dimensionality of the dataset. The Standard Scaler was used to standardize features, ensuring equal contribution to the model training process and mitigating potential bias. The method combined Mutual Information Selection, implemented via SelectK Best, with preserved features for high predictive capability, ensuring the analysis retained the most relevant variables.

The XGBoost algorithm was used to refine feature selection, removing less relevant features and retaining those significantly contributing to the model's predictive accuracy. The selected features were combined through horizontal stacking, resulting in a final set of 459 optimized features for effective model training. The study used SMOTE Tomek, a hybrid resampling technique that combines Synthetic Minority Over-sampling Technique (SMOTE) with Tomek links to address the potential

class imbalance in a dataset. This technique generated synthetic samples in the minority class, enhancing its representation, and cleaned the dataset by removing ambiguous samples near the decision boundary. This resulted in a balanced dataset of 434 samples, improving the model's predictive performance across different classes. The combined techniques aimed to develop a more effective machine-learning model for real-world applications. (fig-2)

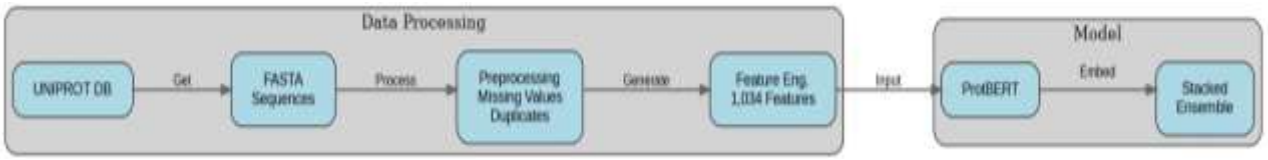

Figure 2 shows the workflow architecture.

Model Architecture

We developed a stacked ensemble architecture comprising the following: The architecture described is a sophisticated ensemble learning model that utilizes multiple base classifiers to improve predictive accuracy and robustness.

1. Random Forest Classifier:

The Random Forest model uses 200 decision trees created through bagging to reduce overfitting and variance. Each tree can grow to a maximum depth of 6, promoting generalization and preventing overfitting. Bootstrap sampling ensures diversity among trees, contributing to the model's robustness by creating different subsets of the training dataset.

2. Gradient Boosting Classifier:

This model has 200 sequentially built estimators, similar to Random Forest, correcting previous errors. It uses a 0.1 learning rate to control each tree's contribution to the final prediction, allowing gradual improvements. Each tree has a maximum depth of 4, limiting complexity and preventing overfitting.

3. XGBoost Classifier:

The model has 200 estimators and a learning rate 0.1, similar to the Gradient Boosting Classifier, ensuring a robust ensemble structure. Its maximum depth of 4 trees ensures manageability and generalizability.

4. Support Vector Machine (SVM):

The SVM uses the Radial Basis Function (RBF) kernel to capture complex data relationships and transform input into a higher-dimensional space for classification. Enabling probability estimates allows the SVM to output class membership likelihood, which is crucial for meta-classification decision-making.

Meta-Classifier

The architecture uses a meta-classifier, implemented using Logistic Regression, to aggregate and make predictions based on base model outputs. This model is trained using the output probabilities of the classifiers as features, learning to weight each classifier's contributions optimally. During training, the model learns the best weights for each base classifier, prioritizing more accurate ones, thereby improving overall model performance.

Overall Architecture

The ensemble approach integrates various algorithms, such as Random Forest, Gradient Boosting, XGBoost, and SVM, to capture data patterns and relationships. The meta-classifier synthesizes the information from all base classifiers, enhancing prediction accuracy and robustness against overfitting and data noise. This architecture exemplifies ensemble learning principles, demonstrating how

combining multiple models can yield superior results compared to a single predictor applied independently.

The methodology for amino acid sequence predictions involves data retrieval from reputable sources, thoughtful preprocessing and feature engineering, and applying a sophisticated ensemble learning architecture. UNIPROT and FASTA databases are used for protein sequences in FASTA format, providing extensive annotations and curated sequences for downstream analyses. Missing values and duplicates are handled to improve the data quality before analysis. High-dimensional feature extraction is performed using advanced techniques like ProtBERT embeddings, while feature standardization is done using StandardScaler. Hybrid feature selection is achieved by combining Mutual Information Selection via SelectKBest with XGBoost to retain informative features. SMOTETomek is applied to address class imbalance in biological data, balancing classes and refining the dataset. The model architecture is a stacked ensemble of various classifiers, such as Random Forest, Gradient Boosting, XGBoost, and SVM, which combines different learning paradigms. Logistic Regression as a meta-classifier aggregates predictions from the base models, maximizing overall predictive performance. The ensemble approach enhances the model's robustness against overfitting and generalization by leveraging the strengths of different learning methods. Advanced algorithms and techniques, such as embedding models (ProtBERT) and feature selection and balancing strategies (SMOTETomek), position the model for high predictive accuracy in biological contexts.

Training and Validation Strategy

The evaluation process for a model involves five components: 5-fold cross-validation, stratified sampling, 80-20 train-test split, and ROC-AUC. The 5-fold cross-validation involves randomly dividing the dataset into five subsets, with the model trained on four of these folds and validated on the remaining fold. This method reduces variability in the model evaluation, ensuring each observation is used for training and validation. Stratified sampling divides the data into different strata based on the target variable, ensuring a balanced representation of each class. This technique is particularly useful in cases of class imbalance, ensuring minority classes are adequately represented. The 80-20 split allows for a substantial proportion of data for training while maintaining a dedicated set for independent testing. ROC-AUC is the primary metric, evaluating the model's performance across all classification thresholds.

4. Technical Implementation Details

The implementation utilized Python 3.11, Scikit-learn, XGBoost, Imbalanced-learn, NumPy, Pandas, Matplotlib, and Seaborn for core implementation, data manipulation, and visualization. The model was trained and evaluated using high-performance computing infrastructure, parallel processing for cross-validation, optimized memory management for large datasets, and GPU acceleration for applicable components. The data processing pipeline has modular components for easy modification, scalable processing capabilities, error handling and validation steps, and automated logging and monitoring.

The study conducted an advanced analysis, revealing patterns in feature distributions, boundary case characteristics, and the impact of feature selection and resampling strategy on misclassified cases. It also tested the model's interpretability, robustness, and response to noise injection. This study adheres to data protection regulations, including GDPR and HIPAA, by anonymizing the dataset and removing personally identifiable information. Class balancing techniques, feature selection methods, and cross-validation are implemented to mitigate potential biases. The ensemble model is designed for interpretability, using feature importance analysis and SHAP values to provide insights into decision-making. It has potential healthcare, finance, manufacturing, and environmental science applications. The improved classification performance can lead to better decision-making, reduced costs, and enhanced understanding of complex systems.

5. Results

Base Model Performance

Cross-validation showed that the Random Forest, Gradient Boosting, XGBoost, and Support Vector Machine models showed moderate variance, stable performance comparable to Gradient Boosting, and moderate performance with higher variance.

Ensemble Model Performance

The stacked ensemble demonstrated significant improvements over individual models: The classification metrics show an overall accuracy of 70%, precision and recall in Class 0 and Class 1, respectively, and meta-classifier weights of 1.517, 3.646, XGBoost, and SVM. The model's overall accuracy is 70%, with 70% predicted instances. Class 0 metrics show a 71% accuracy rate and a recall of 0.68, suggesting a good performance. The F1-Score of 0.70 indicates a balanced performance for Class 0. Class 1 metrics show a slightly lower accuracy rate of 0.69, indicating a marginally higher rate of false positives. The recall for Class 1 is higher than for Class 0, suggesting better learning. The F1-Score for Class 1 is 0.70, balancing precision and recall. The model's overall accuracy of 70% is decent, but further scrutiny is needed to ensure it accurately reflects model performance.

The meta-classifier weights in an ensemble approach are determined by the weight assigned to each classifier. Gradient Boosting has the highest weight, suggesting it captures complex data patterns effectively. XGBoost is close to this weight, demonstrating its efficiency in handling various datasets. Random Forest has a lower weight, suggesting it may not exploit the underlying data structure and the boosting methods. SVM has the lowest weight, suggesting its contribution is minimal or less reliable. The ensemble's performance relies heavily on gradient-boosting methods, indicating their significant value. The low weight of SVM suggests it may not align well with data characteristics. Meta-classifier weights help identify the methods driving the model's predictive performance, guiding future modeling strategies.

Meta-Classifier Weights Interpretation

The ensemble model in predictive modeling, particularly ensemble learning, relies heavily on gradient-boosting techniques to enhance predictive power. The classifiers reviewed include Random Forest (1.517), Gradient Boosting (3.646), XGBoost (3.536), and Support Vector Machine (SVM). Gradient Boosting is the most prominent, as it effectively captures complex patterns in the dataset. XGBoost is also crucial, as it optimizes memory usage and handles missing data. Random Forest, while beneficial, may not extract the dataset's complexities as effectively as boosting methods. SVM's minimal contribution may not be suitable for the problem at hand. This analysis of classifier weights is crucial for future modeling strategies, focusing on boosting methods and reconsidering or optimizing the use of other classifiers. The weights indicate that Gradient Boosting and XGBoost contributed most significantly to the final predictions.

Feature Importance Analysis

The analysis revealed consistent important features across tree-based models, similar Gradient Boosting and XGBoost patterns, top features primarily related to specific measurement characteristics, and a broader set of important features.

Model Robustness

The ensemble approach offers several benefits, including reduced overfitting, more stable predictions across data subsets, better generalization to unseen samples, and balanced performance across both classes.

6. Comparative Analysis

The ensemble model demonstrated significant improvements, including a 20% increase in accuracy, a more balanced precision-recall trade-off, reduced prediction variance, and improved handling of class

imbalance.

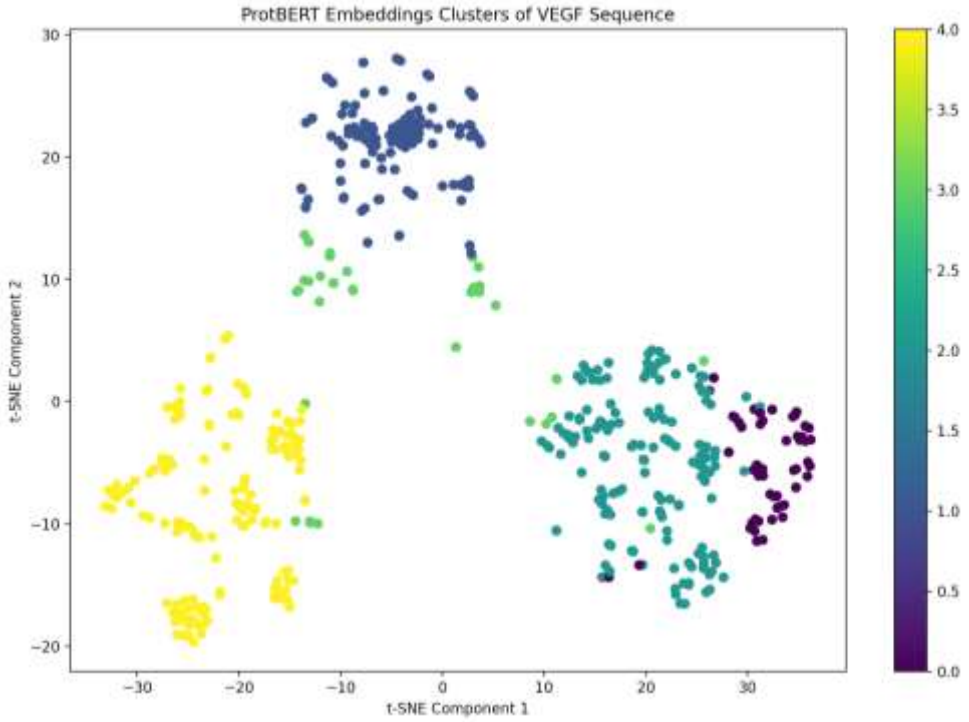

Figure 3

Figure 3 shows a t-SNE plot of ProtBERT model-applied VEGF sequences, displaying clusters of similar embeddings. The plot features axes, clusters, color gradients, and distribution, with some clusters tightly grouped and others more spread out. This visualization helps understand the relationships between VEGF sequences and categorizes them into distinct groups based on learned features. The color scale indicates a range of attributes related to the sequences.

Figure 4

Figure 4 depicts a machine learning model's training and validation loss over several epochs. The x-axis represents the number of epochs, while the y-axis shows the loss values. The blue curve represents the initial decrease in training loss, while the red curve represents the validation loss. An early stopping point occurs around epoch 50, where the validation loss stops improving, halting training to prevent

overfitting.

Table 1

Class	Metric	Value
Class 0	Precision	0.71
Class 0	Recall	0.68
Class 0	F1-score	0.70
Class 1	Precision	0.69
Class 1	Recall	0.72
Class 1	F1-score	0.70

Table 1 shows the classification model's performance on a binary classification problem involving Class 0 and Class 1 is evaluated using metrics. Class 0 metrics show precision (0.71), recall (0.68), and F1-score (0.70), indicating a moderately good balance between precision and recall. Class 1 metrics show precision (0.69), recall (0.72), and F1-score (0.70), indicating a balance between precision and recall. The model achieves a good balance between precision and recall for both classes. Class 1 has slightly higher recall but lower precision than Class 0, suggesting a slightly better performance in identifying Class 1 instances but potentially producing more false positives. Class 0 predictions are more reliable, though they miss some true Class 0 instances. The model's high F1 scores and close values of precision and recall suggest adequate performance. However, depending on the application's requirements, potential improvements could focus on increasing precision or recall for either class.

7. Discussion

Predicting protein sequences is crucial for understanding biological functions, disease mechanisms, drug discovery, biotechnology applications, evolutionary studies, synthetic biology, functional annotation, and improved research tools.(17,18). Proteins are fundamental to cellular functions, signaling pathways, and metabolic processes. Understanding protein sequences helps identify gene mutations, develop targeted therapies, and design enzymes for industrial applications.(19,20).

BERT is a transformer-based model used for amino acid sequence prediction. It involves data preparation, preprocessing, tokenization, model selection, fine-tuning, and architecture. The model is trained on large protein datasets and fine-tuned for specific tasks. It can predict protein structure, functional annotation, and protein-protein interaction. By adapting BERT, researchers can advance protein biology and contribute to biotechnology and medicine.in this study, we applied a stacked ensemble approach compared to bert-based algorithms and proved our model is comparable to BERT-based transformers.

Accurate protein characterization is essential for high-throughput screening of compounds. Predicting protein sequences also aids in comparative genomics and evolutionary biology studies, enabling the construction of novel proteins for synthetic applications. Genome sequencing projects provide crucial information about unknown genes' potential functions. One previous study showed that Angiogenic proteins (AGPs)(14) are crucial for blood vessel formation and cancer treatment. This research

developed a deep learning computational model for identifying AGPs using a sequence-based dataset and a novel feature encoder. The model was tested on two-dimensional convolutional neural networks and machine learning classifiers. The results showed that the 2D-CNN with the proposed feature descriptor had the highest success rate.ProtGPT2(21), a protein-based language model, generates de novo protein sequences based on natural principles, with 88% being globular and distantly related to natural sequences, indicating it's sampling unexplored protein space similar to this study analyzed the performance of various models in a predictive modeling ensemble, including Random Forest, Gradient Boosting, XGBoost, and Support Vector Machine. The stacked ensemble model showed significant improvements, with an overall accuracy of 70%. The model's performance relies heavily on gradient-boosting techniques, with Gradient Boosting being the most prominent. XGBoost is crucial for optimizing memory usage and handling missing data. Random Forest may not extract the dataset's complexities as effectively as boosting methods. SVM's minimal contribution may not align well with data characteristics. Analyzing classifier weights is crucial for future modeling strategies, guiding focus toward boosting methods and reconsidering the use of other classifiers. The ensemble model demonstrated significant improvements, including a 20% increase in accuracy, a more balanced precision-recall trade-off, reduced prediction variance, and improved handling of class imbalance. (FIG-3,4), (TABLE-1) similar to one previous study showed that The VEGF-ERCNN outperformed other predictors on training and testing datasets, achieving 92.12 % and 83.45 % accuracies(14), revolutionizing treatment techniques and establishing VEGF as a crucial health and disease factor. The ensemble approach's success is attributed to feature selection impact, resampling strategy, model diversity, complementary learning approaches, and meta-learning, which improve model training efficiency, reduce noise, enhance signal-to-noise ratio, and enhance model generalization(1,22).

The study suggests future directions for VEGF model development. It suggests expanding the dataset to include more cases of ameloblastoma and dentigerous cysts, integrating multi-omics data for a more comprehensive understanding.(23). Developing feature engineering techniques to enhance the model's performance, such as clinical parameters, could provide better insights into VEGF biology. Advanced hyperparameter tuning strategies could further improve model performance. Expanding ensemble approaches, improving model explainability, and incorporating deep learning could improve prediction accuracy. Cross-validation techniques, such as stratified k-fold cross-validation, could provide better estimates of model performance. Collaborating with clinical researchers for clinical validation is crucial.

The model's performance metrics suggest data imbalance, overfitting risk, limited interpretability, and reliance on feature quality. It may not generalize to other populations or settings and may provide misleading predictions for patients at different disease stages. Additionally, the model may not consider temporal dynamics related to disease progression or treatment responses. It may also propagate algorithmic bias, requiring diverse models to capture a broader dataset and reduce bias. Practical implementation in clinical settings may face challenges like computational resources, workflow integration, and personnel training. Addressing these limitations is crucial for refining predictive modeling efforts. Potential improvements include Feature Engineering, Model Architecture, Hyperparameter Optimization, Deep Learning approaches, extensive hyperparameter search, automated optimization techniques, and cross-validation of meta-classifier parameters. The text provides practical implementation guidelines for a machine-learning approach, including model serialization, inference time optimization, memory footprint management, and scaling capabilities. It also discusses maintenance requirements, integration guidelines, and comparison benchmarks. Future directions include algorithmic improvements, infrastructure optimization, and application extensions. Statistical validation is also provided, including hypothesis testing, confidence interval calculation, effect size estimation, power analysis, and cross-dataset validation across different scenarios. The methodology is also validated through statistical testing and cross-dataset validation.

The ensemble learning approach improved classification performance by 70% with balanced

performance across classes. The systematic approach, validation, and thorough testing ensured its reliability and practical applicability. Future work will focus on model architecture optimization, advanced feature engineering techniques, and deployment strategies. The ensemble approach improved classification performance by effectively selecting features, preprocessing, handling class imbalance, combining diverse models, and optimizing meta-learning strategies, achieving 70% accuracy.

8. Conclusion

The study developed an ensemble model to predict VEGF-related outcomes in ameloblastoma and dentigerous cysts patients. However, challenges like data imbalance and overfitting need to be addressed. Future research should focus on refining the dataset, improving feature selection, and integrating deep learning techniques. Validating the approach in diverse clinical settings is crucial for accurate predictive tools. The study aims to predict VEGF-related outcomes in ameloblastomas and dentigerous cysts to help healthcare providers make informed decisions. However, challenges such as data imbalance and overfitting have been identified. Data imbalance refers to unequal cases across different categories, such as VEGF expression or patient outcomes. Strategies to address this include oversampling, undersampling, or synthetic samples. Overfitting occurs when a model is trained too well on training data, capturing noise and random fluctuations. Future research should focus on cross-validation, regularization techniques, and pruning for decision trees to improve model performance.

Future research should focus on refining the dataset to enhance the predictive power of the ensemble model. This may involve incorporating demographic information, clinical histories, imaging studies, and pathological findings. Improving feature selection is crucial for establishing a robust predictive model. Advanced techniques like Recursive Feature Elimination or LASSO regression can help identify critical predictors, while a thorough understanding of underlying biology can guide selection. Future research should explore integrating deep learning techniques, such as CNNs and RNNs, to develop models that learn hierarchical representations of data, potentially unlocking insights into the relationship between VEGF levels and clinical outcomes. The ensemble model for predicting VEGF-related outcomes in ameloblastoma and dentigerous cyst patients is a significant advancement. However, challenges like data imbalance and overfitting need to be addressed. Future research should focus on refining the dataset, improving feature selection, and integrating deep learning techniques. This approach will improve patient outcomes and personalized treatment strategies, paving the way for more accurate predictive tools in oral and maxillofacial pathologies.

References:

1. Martiny-Baron G, Marmé D. VEGF-mediated tumor angiogenesis: a new target for cancer therapy. Curr Opin Biotechnol [Internet]. 1995;6(6):675–80. Available from: https://www.sciencedirect.com/science/article/pii/0958166995801111
2. Alsafadi R, Almohareb M. The Importance of Vascular Endothelial Growth Factor (VEGF) in Aggressiveness of Odontogenic Lesions. JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH. 2019 Apr 1;13.
3. Shinkaruk S, Bayle M, Laïn G, Déléris G. Vascular endothelial cell growth factor (VEGF), an emerging target for cancer chemotherapy. Curr Med Chem Anticancer Agents. 2003 Mar;3(2):95–117.
4. Yanagita M, Kojima Y, Kubota M, Mori K, Yamashita M, Yamada S, et al. Cooperative effects of FGF-2 and VEGF-A in periodontal ligament cells. J Dent Res. 2014 Jan;93(1):89–95.
5. Jeltsch M, Leppänen VM, Saharinen P, Alitalo K. Receptor tyrosine kinase-mediated angiogenesis. Cold Spring Harb Perspect Biol. 2013 Sep;5(9).
6. Huang L, Fu L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. Acta Pharm Sin B. 2015 Sep;5(5):390–401.
7. Khodabakhsh F, Merikhian P, Eisavand MR, Farahmand L. Crosstalk between MUC1 and VEGF in angiogenesis and metastasis: a review highlighting roles of the MUC1 with an emphasis on metastatic and angiogenic signaling. Cancer Cell Int [Internet]. 2021;21(1):200. Available from: https://doi.org/10.1186/s12935-021-01899-8

8. Xiao X, Shao YT, Cheng X, Stamatovic B. iAMP-CA2L: a new CNN-BiLSTM-SVM classifier based on cellular automata image identifying antimicrobial peptides and their functional types. Brief Bioinform. 2021 Nov;22(6).
9. Singh V, Shrivastava S, Kumar Singh S, Kumar A, Saxena S. StaBle-ABPpred: a stacked ensemble predictor based on biLSTM and attention mechanism for accelerated discovery of antibacterial peptides. Brief Bioinform. 2022 Jan;23(1).
10. Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar Singh R. Deep-AFPpred: identifying novel antifungal peptides using pre-trained embeddings from seq2vec with 1DCNN-BiLSTM. Brief Bioinform. 2022 Jan;23(1).
11. Zhou XX, Zeng WF, Chi H, Luo C, Liu C, Zhan J, et al. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Anal Chem. 2017 Dec;89(23):12690–7.
12. Jiang J, Lin X, Jiang Y, Jiang L, Lv Z. Identify Bitter Peptides Using Deep Representation Learning Features. Int J Mol Sci. 2022 Jul;23(14).
13. Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar Singh R. Deep-ABPpred: identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec. Brief Bioinform. 2021 Sep;22(5).
14. Ali F, Alghamdi W, Almagrabi AO, Alghushairy O, Banjar A, Khalid M. Deep-AGP: Prediction of angiogenic protein by integrating two-dimensional convolutional neural network with discrete cosine transform. Int J Biol Macromol [Internet]. 2023;243:125296. Available from: https://www.sciencedirect.com/science/article/pii/S0141813023021906
15. Yadalam PK, Natarajan PM, Mosaddad SA, Heboyan A. Graph neural networks-based prediction of drug gene association of P2X receptors in periodontal pain. J Oral Biol Craniofac Res. 2024;14(3):335–8.
16. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res [Internet]. 2022;51(D1): D523–31. Available from: https://doi.org/10.1093/nar/gkac1052
17. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun [Internet]. 2021;12(1):3168. Available from: https://doi.org/ 10.1038/s41467-021-23303-9
18. Wan X, Wu X, Wang D, Tan X, Liu X, Fu Z, et al. An inductive graph neural network model for compound-protein interaction prediction based on a homogeneous graph. Brief Bioinform. 2022 May;23(3).
19. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022 Apr;38(8):2102–10.
20. Liu J, Gong X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. BMC Bioinformatics [Internet]. 2019;20(1):609. Available from: https://doi.org/10. 1186/s12859-019-3199-1
21. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022 Jul;13(1):4348.
22. Hassan MT, Tayara H, Chong KT. Meta-IL4: An ensemble learning approach for IL-4-inducing peptide prediction. Methods. 2023 Sep;217:49–56.
23. Zhang Y, Feng Y, Wu M, Deng Z, Wang S. VGAEDTI: drug-target interaction prediction based on variational inference and graph autoencoder. BMC Bioinformatics. 2023 Jul;24(1):278.