Physicochemical-Based Deep Learning for Allergenicity Prediction

Charalambos Chrysostomou(✉)

Eratosthenes Center of Excellence, Limassol, Cyprus
charalambos.chrysostomou@eratosthenes.org.cy

Abstract. Predicting protein allergenicity accurately is crucial for food safety and biopharmaceutical development, yet it remains a significant challenge. This paper introduces a novel deep learning framework for allergenicity prediction, employing a one-dimensional convolutional neural network (CNN). Our approach uniquely represents protein sequences using an extensive set of 611 amino acid physicochemical properties, which are then systematically reduced via Principal Component Analysis (PCA) to derive highly informative features. Evaluated on a comprehensive dataset curated from multiple allergen databases, the model utilising the first three principal components (PCA-3) for encoding demonstrates superior performance. It achieved an accuracy of 97.24%, sensitivity of 96.26%, specificity of 97.67%, a Matthews Correlation Coefficient (MCC) of 0.93, and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.97 on the independent test set. These results underscore the power of leveraging PCA-distilled physicochemical features within a CNN architecture for robust and high-accuracy allergenicity prediction, offering a promising advancement in the field.

Keywords: Allergenic proteins · deep learning · convolutional neural network · protein sequence representation · Principal Component Analysis · bioinformatics · allergy prediction · machine learning

1 Introduction

Allergic diseases are now recognised as a major public health problem that afflicts millions of individuals globally. These conditions can range from mild reactions to severe, potentially fatal anaphylaxis, posing diverse clinical challenges and substantial economic burdens on healthcare systems [26]. Precise identification and characterisation of allergenic proteins are also crucial for a number of important applications, such as the production of hypoallergenic foods, safer biopharmaceuticals, and the design of targeted immunotherapies [14,25].

The prediction of allergenic proteins has traditionally depended on sequence similarity and motif recognition methods, which involve comparing protein sequences with known allergens to identify potential allergenic motifs. These

D. Cafolla et al. (Eds.): AIiH 2025, LNCS 16038, pp. 396–407, 2026.
https://doi.org/10.1007/978-3-032-00652-3_28

methods often fail to generalise, particularly in novel proteins, making the construction of an effective predictor [3] challenging. This restriction highlights the necessity for more sophisticated and informative predictive approaches.

In recent years, computational biology and machine learning have experienced rapid progress with the emergence of deep learning (DL) approaches. These models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models, have respectively shown exceptional potential in extracting complex patterns from biologically relevant data. The application of deep learning methods in computational biology has achieved superior performance compared to traditional methods in a number of areas, including protein structure prediction [1], protein property prediction [6], and B-factor prediction of proteins [23].

This study addresses key challenges in allergenicity prediction prevailing in traditional and existing deep-learning methods. Unlike conventional techniques that rely on sequence similarity and motif recognition, we employ a CNN-based model combined with a novel sequence representation method. This method incorporates multiple amino acid indices to capture various physicochemical properties and uses PCA to improve generalisation. Additionally, our comprehensive dataset, compiled from multiple allergen databases, enhances the model's ability to generalise across diverse protein datasets, aiming to improve sensitivity and specificity in allergenic protein prediction.

Our contributions are threefold: we develop a CNN model specifically tailored for allergenic protein prediction using detailed physicochemical properties of amino acids; we create and utilize a robust and diverse dataset from multiple allergen databases for high-quality, comprehensive training data; and we demonstrate that our approach, especially with PCA-reduced features, achieves high performance in accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), comparing it with established benchmarks. The remainder of this paper is structured as follows: Sect. 2 reviews the current state of the art in allergenic protein classification. Section 3 details the materials and methods, including data description, amino acid indices, and the proposed CNN model architecture. Section 4 presents experimental results, their analysis, and a discussion of their implications, providing a detailed comparison of various encoding schemes. Finally, Sect. 5 concludes the paper by discussing the key findings, limitations, and future research directions.

2 Current State of the Art

Traditional and current state-of-the-art methods for predicting allergenic proteins use sequence similarity and motif-based techniques. These methods involve comparing protein sequences with known allergens to detect potential allergenic motifs. However, they often exhibit limited sensitivity and specificity, particularly with novel proteins not present in existing databases [3]. This limitation affects their accuracy across diverse datasets.

Recent allergenic protein classification advancements include traditional and deep learning methodologies [4,21,22,27]. Despite these advancements, significant challenges persist, especially regarding classification accuracy variation among different protein families and species. Techniques such as AllerCatPro 2.0 [21] and AlgPred 2.0 [27] demonstrate high accuracy but are considerably influenced by the representativeness of their training datasets, complicating predictions for novel allergens or proteins from less-studied species. We collected an extensive dataset from multiple allergen databases to overcome this limitation and validate the proposed model. This novel dataset was created to ensure a diverse and high-quality collection of protein sequences for training and evaluation.

Deep learning models encounter significant challenges in processing protein data, requiring the conversion of protein sequences into numerical representations using techniques such as one-hot encoding, position-specific scoring matrices (PSSMs) [23], embedding layers, and biological sequence-specific language models [17]. These encoding methods are essential for analysing large-scale biological data and identifying complex patterns in protein sequences. The choice of encoding method plays an important role in a model's ability to generalise across datasets and capture different allergenic properties [24]. Furthermore, the complexity of protein encodings requires careful selection and optimisation to avoid overfitting and to ensure that the model generalises effectively from training data. This highlights a fundamental constraint in traditional and deep learning approaches: their performance depends on the quality and representativeness of input data, which can vary widely across different studies and databases.

Although recent advancements in allergenic protein classification using traditional and deep learning methods have enhanced predictive accuracy, challenges regarding accuracy, generalisation, and data encoding remain. Addressing these issues requires continuous refinement of encoding strategies, improved data collection and annotation, and the development of novel models capable of learning from complex biological datasets without compromising their generalisation ability across diverse protein types. This study addresses key challenges in allergenic protein prediction prevalent in traditional and existing deep-learning methods. Unlike conventional techniques that rely on sequence similarity and motif recognition, we employ a CNN-based model combined with a novel sequence representation method. This method incorporates multiple amino acid indices to capture various physicochemical properties and uses PCA to improve generalisation. Additionally, our comprehensive dataset, compiled from multiple allergen databases, enhances the model's ability to generalise across diverse protein datasets, aiming to improve sensitivity and specificity in allergenic protein prediction.

3 Materials and Methods

3.1 Data

For this study, we compiled an extensive dataset from several publicly available allergen databases to create a diverse and high-quality collection of protein

sequences for allergenicity prediction. The sources and the number of protein sequences collected from each are detailed in Table 1. The AlgPred 2.0 dataset [27] was used as a primary source for both allergenic and non-allergenic proteins; we included all 3219 protein sequences listed as allergens and all 10075 sequences listed as non-allergens available at the time of data collection to ensure a comprehensive representation from this benchmark. Other databases were used to augment the set of allergenic proteins.

Table 1. Number of protein sequences collected from various allergen databases. The AlgPred 2.0 entries represent the complete sets of allergens and non-allergens used from that source.

Database	Number of Protein Sequences
AlgPred 2.0 (Allergens)	3219
AllergenOnline v22	2286
Allergen.org	942
Allergome (Reviewed)	1310
COMPARE	2748
UniProtKB	948
Unique Allergenic Total	**4336**
AlgPred 2.0 (Non-Allergens)	10075

After removing duplicates from the combined allergenic sources, the dataset included 4336 unique allergenic protein sequences. This ensures a broad and diverse representation for robust model training. The 10075 non-allergenic protein sequences from AlgPred 2.0 establish a reliable non-allergenic benchmark, crucial for training and validating models to accurately differentiate between allergenic and non-allergenic proteins. The data was split into training (80%), validation (10%), and testing (10%) sets, and results were validated using 10-fold cross-validation on the training/validation portion during model development.

3.2 Amino Acid Indices

Proteins are composed of amino acids, each possessing distinct physicochemical and biological attributes. These properties are crucial for analysing protein sequences, particularly when predicting allergenicity. In bioinformatics, specific amino acid indices numerically represent these attributes, ranging from structural to thermodynamic and evolutionary properties.

The AAindex database [16] serves as an extensive repository of amino acid indices, capturing the varied nature of amino acids. This database facilitates the transformation of protein sequences from alphabetical to numerical formats, which is essential for computational analysis, especially with machine learning algorithms. From the AAindex database, 528 indices were identified. From a

literature review of recent papers, 83 additional amino acid indices were identified [2,12,13,15,18,28]. In total, 611 amino acid indices were used in this study.

Using 611 scales to encode proteins poses a significant challenge due to the high dimensionality ("curse of dimensionality"), which can lead to computational inefficiencies, increased risk of overfitting, and difficulties in model training. Previous work [5,7–11] has demonstrated that using certain physicochemical properties, such as hydrophobicity, can be effective in classifying protein sequences. However, by utilising a broader range of amino acid indices and applying PCA [20], our approach aims to capture a more comprehensive set of properties, potentially leading to a more accurate and robust classification of allergenic proteins. These indices capture various physicochemical properties of amino acids, such as hydrophobicity, charge, and molecular weight, offering a detailed numerical representation of amino acids.

Principal Component Analysis (PCA) is used to reduce the dimensionality of the 611 amino acid indices. This is accomplished by identifying orthogonal principal components that capture the directions of maximal variance within the data. This transformation retains essential information whilst reducing data redundancy, leading to a more compact and computationally efficient representation. The principal components created can then be used as features for amino acid encoding in predictive deep-learning models. Focusing on the components that explain the majority of variance minimises the influence of noise and redundant information, potentially improving downstream models' performance and computational efficiency. These derived features effectively reveal underlying patterns and relationships within the amino acid dataset, which may be particularly relevant for predicting allergenic potential. A comprehensive list of the amino acid indices utilised in the PCA-based encodings is provided in Table 2.

Given the dataset's composition of 20 unique amino acids described by 611 indices, the maximum number of principal components that can be extracted is 19 (number of amino acids - 1). This limit is imposed by the lower dimension of the original data matrix (20 amino acids vs. 611 indices). This dimensionality reduction improves model interpretability, making PCA a valuable preprocessing step. Due to the specific dataset structure and the nature of PCA, the final principal component (PC20, if calculated from a 20xN matrix) would have zero variance and is thus excluded, leaving 19 informative components.

3.3 Proposed Method

This study employed a 1D convolutional neural network (CNN) to predict protein allergenicity. The choice of a 1D CNN was motivated by its effectiveness in capturing local patterns and motifs within sequential data, such as protein sequences represented by physicochemical properties. While architectures like LSTMs can model long-range dependencies and handle variable-length sequences directly, 1D CNNs are computationally efficient and have demonstrated strong performance in various bioinformatics sequence classification tasks, often excelling when discriminative local features are key. Our fixed-length input approach (detailed below) makes CNNs a suitable choice. The CNN

architecture is depicted in Fig. 1 and comprises three main convolutional blocks followed by fully connected layers.

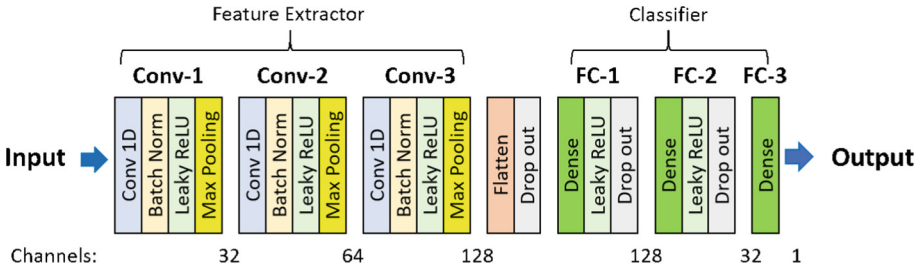

Fig. 1. Architecture of the 1D Convolutional Neural Network (CNN) for Allergenic Protein Prediction.

The network structure comprises three convolutional blocks, each containing a 1D convolutional layer, Batch Normalization, a Leaky ReLU activation function, and Max Pooling. The first block uses 32 filters with a kernel size of 3, the second employs 64 filters with a kernel size of 3, and the third utilizes 128 filters with a kernel size of 3. Batch Normalization is incorporated to stabilize and speed up training, while Leaky ReLU addresses the vanishing gradient problem, and Max Pooling with a pool size of 2 reduces dimensionality and offers some translational invariance. Following the convolutional blocks, the feature maps are flattened and then processed by two fully connected dense layers with 128 and 32 units, respectively. Both dense layers use Leaky ReLU activation and apply Dropout with a rate of 0.5 for regularization to mitigate overfitting. The final output layer consists of 2 units and employs a Softmax activation function to produce a probability distribution over the two classes (allergenic and non-allergenic).

The input to the model consists of protein amino acid sequences. These sequences are numerically encoded using one of the schemes detailed in Sect. 3.2 (e.g., Integer, Hydrophobicity, or PCA-based). This transforms each amino acid into a numerical vector of length N_f, where N_f is the number of features for the chosen encoding scheme (e.g., $N_f = 3$ for PCA-3). To handle sequences of varying lengths, all sequences were padded with zeros to a fixed length of 3416, which was determined as the maximum sequence length observed in our compiled dataset. While padding shorter sequences can introduce artificial signals, it is a common technique for batch processing in CNNs; the use of Focal Loss (described below) and an extensive dataset aims to mitigate potential adverse effects on learning meaningful patterns. Input data was also normalised to ensure feature values fall within a consistent range, aiding training convergence. Thus, each protein is represented as a matrix of size $3416 \times N_f$.

Initially, Binary Cross-Entropy (BCE) loss was considered. However, given the imbalance in our dataset (more non-allergenic sequences than allergenic ones,

Table 2. Amino acid encoding schemes used in the study, including integer, hydrophobicity, and principal components (PC1 to PC19). Values for PCs are illustrative of the transformed scale for each amino acid.

Encoding	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Integer	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
Hydrophobicity	0.6	0.6	0.1	0.5	1.1	0.5	0.0	0.1	0.6	2.2	1.5	1.1	1.2	2.0	2.0	0.1	0.1	2.7	1.9	1.3
PC1	0.9	−11.0	−16.4	−20.2	10.2	−14.1	−9.4	−15.0	−1.1	22.3	18.9	−14.3	17.1	21.0	−16.2	−12.7	−4.8	18.6	8.6	17.5
PC2	4.1	−17.4	1.1	−1.5	8.2	−12.1	−9.1	22.4	−7.9	4.5	1.8	−14.2	−5.8	−0.8	15.6	8.8	6.0	−7.8	−2.6	6.8
PC3	17.3	−2.8	−4.5	−0.3	−8.8	10.6	0.9	2.6	−6.1	3.8	13.4	5.0	2.5	−3.5	−11.8	1.6	0.5	−14.1	−13.5	7.3
PC4	−0.2	2.1	−5.5	−2.6	−16.1	1.2	−0.3	−5.1	−4.9	3.9	5.9	2.5	−2.7	2.7	19.3	−3.4	−1.3	1.5	2.1	1.0
PC5	−4.0	11.2	2.6	−8.2	−6.8	−10.2	−1.5	6.9	−1.4	3.2	1.3	5.9	−6.8	−0.5	−7.6	3.9	4.1	−3.1	6.4	4.7
PC6	−0.6	7.0	−4.1	−4.8	11.0	−1.9	2.5	−9.8	0.5	2.0	−1.6	0.7	−0.9	−4.6	6.2	1.7	4.1	−8.4	−4.3	5.4
PC7	−3.8	−1.3	0.4	9.4	−0.9	5.0	−0.6	−5.6	−5.9	4.3	−0.9	−5.1	−7.7	−0.1	−3.8	2.1	4.2	−0.5	5.3	5.5
PC8	−1.2	−6.6	5.6	−0.4	−5.2	−3.3	1.1	−5.8	8.7	0.6	0.7	−0.7	2.1	2.0	−0.1	4.2	3.2	−5.7	0.8	0.0
PC9	3.9	−1.7	−0.9	−2.5	−0.5	−2.4	1.5	−4.0	−3.9	−5.0	1.7	0.4	−1.5	−1.6	−0.8	7.4	6.6	8.1	−2.1	−2.5
PC10	−2.2	−7.0	3.2	−1.0	4.2	−0.4	0.1	−1.6	−2.4	1.8	3.1	10.4	−4.3	−1.9	0.4	−0.9	−3.6	0.7	1.2	0.1
PC11	1.1	4.3	2.9	2.1	2.6	−0.8	−6.0	−2.0	2.5	−3.0	8.4	−2.7	−5.1	2.9	0.5	0.9	−3.2	−0.9	−1.3	−3.3
PC12	2.4	−1.3	−4.1	−0.7	−0.9	2.0	−3.7	1.0	8.6	0.4	−2.6	1.2	−6.1	−2.5	0.0	−1.0	0.8	4.0	−1.5	4.1
PC13	1.3	−2.0	−4.8	−3.8	1.6	3.9	2.1	0.9	0.8	−5.4	1.5	−0.5	−1.6	1.6	−0.2	1.1	−0.1	−3.8	8.4	−1.2
PC14	0.5	−0.1	−5.8	7.1	−0.2	−5.6	−2.1	−0.3	0.6	−1.7	1.1	3.3	3.1	−1.8	−0.2	0.9	−0.2	−1.1	2.1	0.2
PC15	0.1	−0.3	−1.1	0.3	0.7	0.7	−3.7	−0.1	−1.5	−0.9	−3.5	3.8	−0.6	8.3	−0.1	−1.2	4.2	−1.7	−2.2	−1.1
PC16	6.0	0.7	3.5	−1.0	−0.1	0.2	−5.4	−2.4	−1.7	0.5	−3.8	0.4	1.9	−2.0	0.6	0.2	−1.5	−0.4	4.0	0.3
PC17	5.1	0.0	0.5	1.6	0.0	−4.1	5.6	−0.8	−0.2	−0.3	−2.1	−0.7	−3.8	3.2	−0.1	−1.6	−3.0	−0.2	−0.1	1.1
PC18	−1.5	0.3	−1.2	−0.4	−0.4	1.0	−0.5	−0.2	−0.4	−0.1	−1.8	0.4	0.6	2.1	0.0	6.3	−5.2	0.3	−1.2	2.0
PC19	−1.1	−0.2	2.2	0.1	−0.5	−0.1	−0.3	−0.1	−0.7	−5.5	0.6	0.1	1.2	0.2	0.5	−1.9	0.0	0.5	−0.5	5.7

see Table 1), which can bias the model towards the majority class, we adopted Focal Loss [19]. Focal Loss modifies the standard cross-entropy loss by adding a modulating factor that down-weights the contribution of well-classified examples (typically from the majority class). This allows the model to focus more on hard-to-classify examples, which often belong to the minority class, thereby improving performance on imbalanced datasets.

4 Results and Discussion

In this section, we present the performance evaluation of our proposed CNN model using various encoding schemes for allergenic protein prediction. The evaluation metrics employed include accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The performance of the encoding schemes on the training, validation, and testing sets is presented in Tables 3. Table 3 provides a detailed comparison on the unseen testing set, including a comparison with the reported metrics of AlgPred 2.0 [27] for context, as it is a widely recognised tool in allergenicity prediction.

On the testing set (Table 3), the Integer encoding achieved an accuracy of 93.21%, while Hydrophobicity encoding performed better at 96.79%. The PCA-based encodings demonstrated varied performance: PCA-3 yielded the highest accuracy at 97.24%, followed by PCA-5 (96.98%), PCA-1 (95.13%), PCA-10 (92.84%), and PCA-19 (92.60%).

Sensitivity, the ability to correctly identify allergenic proteins, was 81.45% for Integer encoding and significantly higher for Hydrophobicity encoding (93.95%). PCA-3 achieved the best sensitivity among all methods (96.26%), followed by PCA-5 (95.01%). Specificity, the ability to correctly identify non-allergenic proteins, was high for Integer encoding (98.26%). Hydrophobicity (98.01%) and PCA-3 (97.67%) also showed excellent specificity.

The Matthews Correlation Coefficient (MCC), a balanced measure of classification quality, was 0.84 for Integer encoding and 0.92 for Hydrophobicity. PCA-3 and PCA-5 achieved the highest MCC of 0.93. The Area Under the ROC Curve (AUC-ROC) indicates the model's discriminative ability. Integer encoding had an AUC of 0.90, while Hydrophobicity reached 0.96. PCA-3 and PCA-5 achieved the highest AUC of 0.97 among our tested encodings.

Analysis of PCA-3 Performance: The PCA-3 encoding, which utilises the first three principal components derived from the 611 physicochemical properties, consistently demonstrated superior or near-best performance across most metrics. This suggests that these three components capture the most critical information relevant to allergenicity from the high-dimensional physicochemical space. While the exact interpretation of individual principal components can be complex, generally:

– **PC1** often represents the most dominant source of variation, which in amino acid properties frequently relates to hydrophobicity/hydrophilicity and size.

Table 3. Performance Comparison. Top: Training & Validation Sets (Mean ± Std. Dev., 10-fold CV). Bottom: Testing Set (Mean ± Std. Dev., 10-fold CV test splits) & AlgPred 2.0.

Training and Validation Sets

Set	Metric	Integer	Hydrophobicity	PCA-1	PCA-3	PCA-5	PCA-10	PCA-19
Training	Acc (%)	91.2±1.38	96.08±0.88	94.89±0.99	97.06±0.51	96.49±0.68	91.30±1.32	95.37±2.44
	Sens (%)	83.66±2.69	94.13±1.43	91.60±1.51	96.36±0.78	95.10±1.01	86.49±2.20	95.07±4.16
	Spec (%)	98.67±0.35	98.02±0.38	98.16±0.51	97.77±0.27	97.88±1.58	96.11±0.50	95.66±1.64
	MCC	0.83±0.03	0.92±0.02	0.90±0.02	0.94±0.01	0.93±0.01	0.83±0.02	0.91±0.05
	AUC	0.91±0.01	0.96±0.01	0.95±0.01	0.97±0.01	0.96±0.01	0.91±0.01	0.95±0.02
Validation	Acc (%)	92.89±3.93	96.59±3.31	95.68±2.76	97.06±2.30	96.94±2.68	92.94±0.85	91.65±2.34
	Sens (%)	80.95±9.56	93.30±7.15	90.30±5.65	95.80±4.31	94.88±5.22	85.59±1.33	88.61±3.71
	Spec (%)	98.02±1.56	98.00±1.69	97.99±1.54	97.61±1.58	97.84±1.62	96.09±0.81	92.95±2.23
	MCC	0.83±0.10	0.92±0.08	0.90±0.07	0.93±0.06	0.93±0.06	0.83±0.02	0.80±0.05
	AUC	0.89±0.06	0.96±0.04	0.94±0.04	0.97±0.03	0.97±0.03	0.91±0.01	0.91±0.02

Testing Set and AlgPred 2.0 Comparison

Set	Metric	Integer	Hydrophobicity	PCA-1	PCA-3	PCA-5	PCA-10	PCA-19	AlgPred 2.0 [27]
Testing	Acc (%)	93.21±3.84	96.79±3.31	95.13±3.09	**97.24±2.27**	96.98±2.65	92.84±1.15	92.60±2.73	94.23
	Sens (%)	81.45±9.24	93.95±6.97	89.38±6.67	**96.26±3.88**	95.01±4.90	85.73±2.87	89.95±5.92	93.10
	Spec (%)	98.26±1.60	98.01±1.75	97.61±1.59	**97.67±1.62**	97.83±1.72	95.90±0.83	93.73±2.36	95.36
	MCC	0.84±0.10	0.92±0.08	0.88±0.08	**0.93±0.05**	**0.93±0.06**	0.83±0.03	0.83±0.06	0.88
	AUC	0.90±0.05	0.96±0.04	0.93±0.04	**0.97±0.03**	**0.97±0.03**	0.91±0.02	0.92±0.03	0.99

– **PC2 and PC3** capture subsequent orthogonal dimensions of variance, potentially representing properties like charge, polarity, and secondary structure propensity, which are not fully encapsulated by PC1.

The effectiveness of PCA-3 implies that a compact representation focusing on these primary axes of physicochemical variation is more beneficial for the CNN model than using a single property (like hydrophobicity alone), many more PCs (which might introduce noise or redundant information for this specific task), or simple integer encoding. The dimensionality reduction helps in creating a more robust and generalizable feature set, reducing the risk of overfitting to the extensive original 611 properties. The CNN can then effectively learn discriminative patterns from these condensed yet informative features.

Comparison with AlgPred 2.0: Our PCA-3 model achieved an accuracy of 97.24%, sensitivity of 96.26%, specificity of 97.67%, and MCC of 0.93 on our diverse test set. AlgPred 2.0 reported an accuracy of 94.23%, sensitivity of 93.10%, specificity of 95.36%, and MCC of 0.88 [27]. While direct comparisons should be made cautiously due to potential differences in datasets and evaluation protocols (our dataset is a newly compiled one, though it incorporates AlgPred 2.0 data), our model with PCA-3 encoding shows competitive and, on several metrics, improved performance. AlgPred 2.0 reported a higher AUC (0.99); our PCA-3 model achieved an AUC of 0.97. This difference could be attributed to various factors, including the specific composition of the datasets used for the AUC calculation or the inherent strengths of the support vector machine (SVM) based methods used in AlgPred 2.0 for this particular metric. However, the strong performance of our CNN with PCA-3 across accuracy, sensitivity, specificity, and MCC highlights its robustness.

These results indicate that the PCA-3 encoding scheme provides a powerful representation of amino acid properties for allergenicity prediction with our CNN architecture. It achieves a strong balance between sensitivity and specificity, crucial for reliable allergen identification. While other machine learning models (e.g., LSTMs, Transformers, or other SVM configurations) were not explored in this specific study, their investigation could be a subject for future work to further benchmark the PCA-based physicochemical feature representation.

5 Conclusion

This study demonstrates the effectiveness of a physicochemical-based deep learning approach for accurate allergenicity prediction. By encoding protein sequences based on a comprehensive set of amino acid properties reduced by PCA and employing a 1D CNN, our model advances beyond traditional methods, achieving high performance, particularly with the PCA-3 encoding scheme. This work highlights the significance of incorporating detailed physicochemical information in a condensed form to improve the accuracy and robustness of allergenicity prediction. Our findings show that the PCA-3 encoding yields strong results across multiple metrics (Accuracy: 97.24%, Sensitivity: 96.26%, Specificity: 97.67%,

MCC: 0.93, AUC: 0.97), outperforming other tested encoding schemes and showing competitive performance against established benchmarks like AlgPred 2.0. The use of Focal Loss helped address class imbalance, and the comprehensive dataset aids generalizability. While our results are strong, architectures like LSTMs or Transformers that can natively handle variable-length sequences could be explored. Furthermore, this study focused on a specific CNN architecture; other deep learning models or even different machine learning paradigms (e.g., advanced SVMs, ensemble methods) using the proposed PCA-based features could yield further insights. Future directions include exploring more sophisticated encoding strategies, potentially dynamic or attention-based weighting of physicochemical properties.

Acknowledgment. The authors acknowledge the 'EXCELSIOR': ERATOSTHENES: EXcellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (www.excelsior2020.eu). The 'EXCELSIOR' project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

References

1. Abramson, J., et al.: Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 1–3 (2024)
2. Atchley, W.R., Zhao, J., Fernandes, A.D., Drüke, T.: Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. **102**(18), 6395–6400 (2005)
3. Baek, M., Baker, D.: Deep learning and protein structure modeling. Nat. Methods **19**(1), 13–14 (2022)
4. Breiteneder, H., et al.: Biomarkers for diagnosis and prediction of therapy responses in allergic diseases and asthma. Allergy **75**(12), 3039–3068 (2020)
5. Carmona, C.J., Chrysostomou, C., Seker, H., del Jesus, M.J.: Fuzzy rules for describing subgroups from influenza a virus using a multi-objective evolutionary algorithm. Appl. Soft Comput. **13**(8), 3439–3448 (2013)
6. Chandra, A., Tünnermann, L., Löfstedt, T., Gratz, R.: Transformer-based deep learning for predicting protein properties in the life sciences. Elife **12**, e82819 (2023)
7. Chrysostomou, C.: Deep learning-assisted classification of allergenic proteins: an exploration of amino acid indices. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 4881–4883. IEEE (2023)
8. Chrysostomou, C.: Deep learning-based phylogenetic analysis of influenza protein sequences: a Siamese neural network approach. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 4884–4886. IEEE (2023)
9. Chrysostomou, C., Alexandrou, F., Nicolaou, M.A., Seker, H.: Classification of influenza hemagglutinin protein sequences using convolutional neural networks. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1682–1685. IEEE (2021)

10. Chrysostomou, C., Seker, H.: Prediction of protein allergenicity based on signal-processing bioinformatics approach. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 808–811. IEEE (2014)
11. Chrysostomou, C., Seker, H., Aydin, N.: CISAPS: complex informational spectrum for the analysis of protein sequences. Adv. Bioinf. **2015** (2015)
12. Fernández, L., Caballero, J., Abreu, J.I., Fernández, M.: Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene V protein mutants. Proteins Struct. Funct. Bioinf. **67**(4), 834–852 (2007)
13. Gasteiger, E., et al.: Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (ed.) The Proteomics Protocols Handbook. Springer Protocols Handbooks. Humana Press (2005). https://doi.org/10.1385/1-59259-890-0:571
14. EFSA Panel on Genetically Modified Organisms (GMO), et al.: Scientific opinion on development needs for the allergenicity and protein safety assessment of food and feed products derived from biotechnology. EFSA J. **20**(1), e07044 (2022)
15. Huang, J., Kawashima, S., Kanehisa, M.: New amino acid indices based on residue network topology. Genome Inform. **18**, 152–161 (2007)
16. Kawashima, S., Kanehisa, M.: AAindex: amino acid index database. Nucleic Acids Res. **28**(1), 374 (2000)
17. Ko, C.W., Huh, J., Park, J.W.: Deep learning program to predict protein functions based on sequence information. MethodsX **9**, 101622 (2022)
18. Kurgan, L.A., Stach, W., Ruan, J.: Novel scales based on hydrophobicity indices for secondary protein structure. J. Theor. Biol. **248**(2), 354–366 (2007)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
20. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (PCA). Comput. Geosci. **19**(3), 303–342 (1993)
21. Maurer-Stroh, S., et al.: Allercatpro—prediction of protein allergenicity potential from the protein sequence. Bioinformatics **35**(17), 3020–3027 (2019)
22. Nedyalkova, M., Vasighi, M., Azmoon, A., Naneva, L., Simeonov, V.: Sequence-based prediction of plant allergenic proteins: machine learning classification approach. ACS Omega **8**(4), 3698–3704 (2023)
23. Pandey, A., Liu, E., Graham, J., Chen, W., Keten, S.: B-factor prediction in proteins using a sequence-based deep learning model. Patterns **4**(9) (2023)
24. Peng, Z., Wang, W., Han, R., Zhang, F., Yang, J.: Protein structure prediction in the deep learning era. Curr. Opin. Struct. Biol. **77**, 102495 (2022)
25. Pfaar, O., Creticos, P.S., Kleine-Tebbe, J., Canonica, G.W., Palomares, O., Schülke, S.: One hundred ten years of allergen immunotherapy: a broad look into the future. J. Allergy Clin. Immunol. Pract. **9**(5), 1791–1803 (2021)
26. Reznek, L.: The Nature of Disease. Routledge (2022)
27. Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., Raghava, G.P.: AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IGE epitopes. Briefings Bioinf. **22**(4), bbaa294 (2021)
28. Zviling, M., Leonov, H., Arkin, I.T.: Genetic algorithm-based optimization of hydrophobicity tables. Bioinformatics **21**(11), 2651–2656 (2005)