Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

Classification of Cancer Epitope Mutations Using Random Forest, SVM and Sequence-Derived Features

Miftahurrahma Rosyda[1, *], Dinan Yulianto[2]

Faculty of Industrial Technology, Informatics Program, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: [1,*]miftahurrahma.rosyda@tif.uad.ac.id, [2]dinan.yulianto@tif.uad.ac.id
Correspondence Author Email: miftahurrahma.rosyda@tif.uad.ac.id
Submitted: **21/05/2026**; Accepted: **22/06/2026**; Published**: 23/06/2026**

Abstract−Cancer remains a major global health challenge, with increasing incidence and mortality rates worldwide. One promising approach in cancer immunotherapy is the identification of epitopes, particularly mutated epitopes that play a critical role in immune recognition. This study aims to classify cancer epitope mutations using machine learning approaches based on sequence-derived physicochemical features. A dataset consisting of 234 samples was used, comprising into tumor and mutated tumor epitopes. Feature extraction was performed using physicochemical properties such as molecular weight, isoelectric point, aliphatic index, aromaticity, and hydrophobicity. Two machine learning models, namely Support Vector Machine and Random Forest were selected due to their proven effectiveness in biological sequence classification tasks and their robustness in handling small-to-medium-sized datasets. The results show that Random Forest achieved the best performance with an accuracy of 83% and a macro average F1-score of 0.70, while consistently outperforming the Support Vector Machine model across all data partition scenarios. However, further analysis revealed that the model exhibits limitations in detecting mutated epitopes, as indicated by a relatively high false-negative rate. This issue is likely due to the use of global sequence-derived features, which may not effectively capture local variations caused by mutations. This study contributes to the field of computational immunology by providing a comparative evaluation of Random Forest and Support Vector Machine for mutation epitope classification using sequence-derived physicochemical features. In addition, the integration of machine learning analysis with structural bioinformatics interpretation offers further biological insight into mutation-associated epitopes and their potential relevance in cancer immunotherapy.

Keywords: Epitope; Cancer; Machine Learning; Random Forest; Support Vector Machine

1. INTRODUCTION

According to the Indonesian Central Statistics Agency (BPS), Indonesia's population is projected to reach approximately 288.3 million in 2025, accompanied by improvements in health indicators such as increased life expectancy [1]. While these demographic changes reflect positive developments in public health, they also present new challenges, particularly the growing burden of non-communicable diseases. Among these diseases, cancer remains one of the leading causes of death worldwide and poses a significant barrier to improving global health systems. The World Health Organization estimates that 20 million new cancer cases and approximately 10 million deaths are reported annually, with a continuous upward trend [2]. This increasing burden highlights the urgent need for innovative and effective approaches in cancer prevention, diagnosis, and treatment.

In Indonesia, the rising incidence of cancer requires comprehensive strategies that integrate both preventive and therapeutic efforts. Cancer immunotherapy has emerged as a promising approach, particularly through the identification of immunologically relevant epitopes[3]. An epitope is a specific region of an antigen that can be recognized by the immune system and bind to antibodies or T-cell receptors [4]. In the context of cancer, epitopes derived from tumor-associated antigens play a crucial role in activating immune responses against malignant cells [5]. Consequently, accurate identification and prediction of epitopes are essential for developing effective cancer vaccines and targeted immunotherapy strategies.

A major challenge in cancer immunotherapy arises from mutations occurring within epitope sequences. These mutations can alter the structural and physicochemical properties of epitopes, potentially affecting their recognition by the immune system [6]. Even a single amino acid substitution may significantly influence immune binding affinity and antigen presentation. Therefore, analyzing epitope sequences is critical for distinguishing between normal tumor epitopes and mutated tumor epitopes, as well as for improving the accuracy of computational prediction methods.

Traditional techniques for identifying protein structures and epitopes, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR), provide highly accurate results but are often time-consuming, labor-intensive, and costly [7]. These limitations make them less suitable for large-scale analysis of protein datasets. As a result, bioinformatics approaches have become widely adopted as efficient alternatives, enabling large-scale analysis of biological data using computational tools and publicly available databases.

Recent advances in machine learning (ML) have significantly expanded its application in bioinformatics, including epitope-related studies. Algorithms such as Support Vector Machine (SVM) and Random Forest (RF) are capable of learning complex patterns from biological data and have demonstrated strong performance in classification problems [8]. These methods typically rely on sequence-derived features, including physicochemical properties such as molecular weight, isoelectric point, aliphatic index, aromaticity, and hydrophobicity, to represent peptide sequences in numerical form [8]. Such features allow machine learning models to process biological sequences effectively and efficiently.

Machine learning methods offer several advantages for biological data analysis. Their ability to handle high-dimensional and heterogeneous datasets makes them particularly suitable for bioinformatics applications. Ensemble

Copyright © 2026 **Author**, Page 363
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

approaches such as Random Forest further enhance predictive performance by combining multiple decision trees into a robust classification framework. Compared to deep learning approaches, traditional machine learning models are often more appropriate for relatively small datasets, as they require less computational power and offer better interpretability, which is particularly important in biomedical research especially epitope classification [9].

Despite these strengths, several challenges remain in applying machine learning to epitope prediction, especially in the context of mutation classification. One of the main limitations is that sequence-derived physicochemical features generally represent global properties of peptide sequences rather than local variations at specific amino acid positions. As a result, point mutations that occur at specific residues may not significantly change the overall feature values, making mutated epitopes difficult to distinguish from non-mutated ones. This limitation can lead to misclassification, particularly false negatives, where mutated epitopes are incorrectly identified as non-mutated tumor epitopes.

The complexity of biological datasets further complicates this task. Variations in sequence composition, mutation patterns, and class distributions can affect model performance. Moreover, imbalanced datasets may introduce prediction bias toward the majority class. Appropriate preprocessing procedures, feature selection techniques, and rigorous evaluation strategies are therefore required to improve classification reliability and predictive accuracy [10].

Several studies have demonstrated the effectiveness of machine learning approaches for epitope and antigen classification. The LBCPred study compared Random Forest, Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Multilayer Perceptron (MLP) for linear B-cell epitope prediction and reported that Random Forest achieved the highest classification accuracy among the evaluated models[11].Another study showed that SVM produced promising results in tumor T-cell antigen classification using biological sequence information and natural language processing (NLP)-based feature representations [12].EpiGPTope further demonstrated the potential of machine learning for epitope prediction using epitope-like sequence datasets, highlighting the capability of computational approaches in identifying immunologically relevant peptide sequences [13]. DeepBCE introduced a deep learning framework for B-cell epitope prediction in parasitic pathogens and achieved competitive predictive performance[14]. However, deep learning models generally require substantially larger datasets and higher computational resources than conventional machine learning algorithms.

Beyond classification tasks, machine learning and bioinformatics techniques have also been integrated in cancer vaccine design. For example, an integrated computational framework combining machine learning-based epitope screening with structural bioinformatics and AlphaFold modeling was successfully applied to identify potential vaccine candidates against prostate cancer[15]. This study highlights the growing importance of combining predictive models with biological interpretation to support cancer immunotherapy research.

Despite these advances, most previous studies have focused on epitope prediction, immunogenicity assessment, vaccine candidate discovery, or tumor antigen identification. Comparatively little attention has been given to the classification of tumor epitopes and mutation-associated epitopes using physicochemical sequence descriptors.. Therefore, this study evaluates the effectiveness of Random Forest (RF) and Support Vector Machine (SVM) algorithms in distinguishing tumor epitopes from mutated tumor epitopes based on sequence-derived physicochemical features.

Furthermore, the study investigates the limitations of global physicochemical descriptors in capturing mutation-specific characteristics and complements the machine learning analysis with structural bioinformatics interpretation. The main contributions of this research are threefold: (1) evaluating the discriminative capability of physicochemical features for mutation-associated epitope classification; (2) providing a comparative assessment of Random Forest and Support Vector Machine models under multiple train–test split scenarios; and (3) integrating machine learning classification results with structural analyses of ACTN4-derived epitopes to generate additional biological insights. The findings are expected to contribute to the advancement of computational epitope analysis and support the development of more effective cancer immunotherapy strategies.

2. RESEARCH METHODOLOGY

The overall research workflow adopted in this study is illustrated in Figure 1. The study began with dataset collection from the CAPED database, followed by data cleaning to ensure data quality and consistency. Subsequently, sequence-derived physicochemical features were extracted using the BioGEM platform, generating numerical descriptors that represent the biochemical characteristics of each epitope. The extracted features were then used as input for two machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest. Model performance was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis under different train-test split scenarios. Finally, bioinformatics interpretation was performed through molecular modeling and structural analysis to provide additional biological insights into mutation-associated epitopes and support the interpretation of the machine learning results. The systematic workflow presented in this study was designed to ensure the reproducibility and transparency of the analytical process, enabling each stage of the proposed approach to be clearly described and evaluated.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

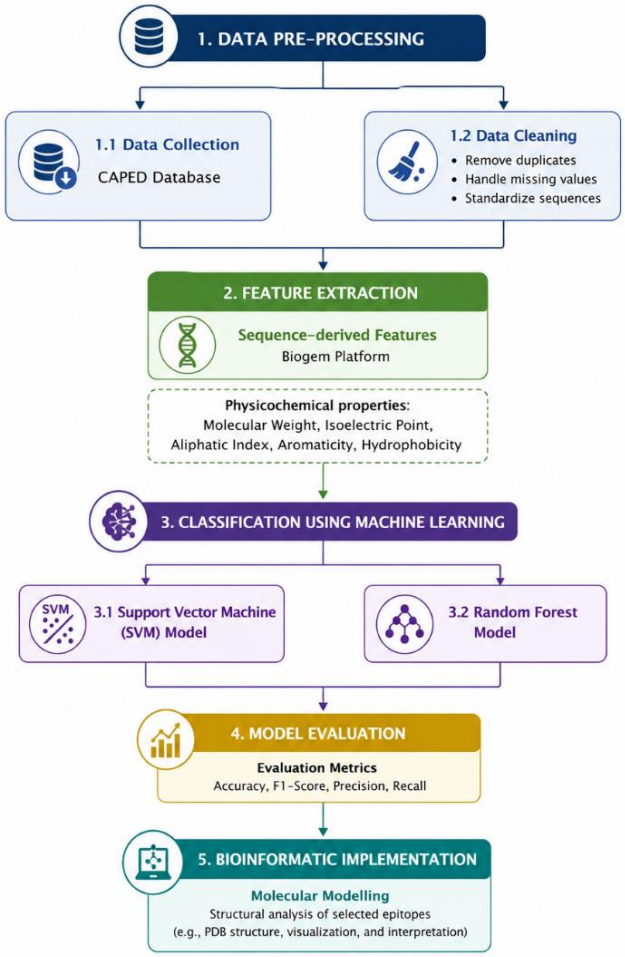
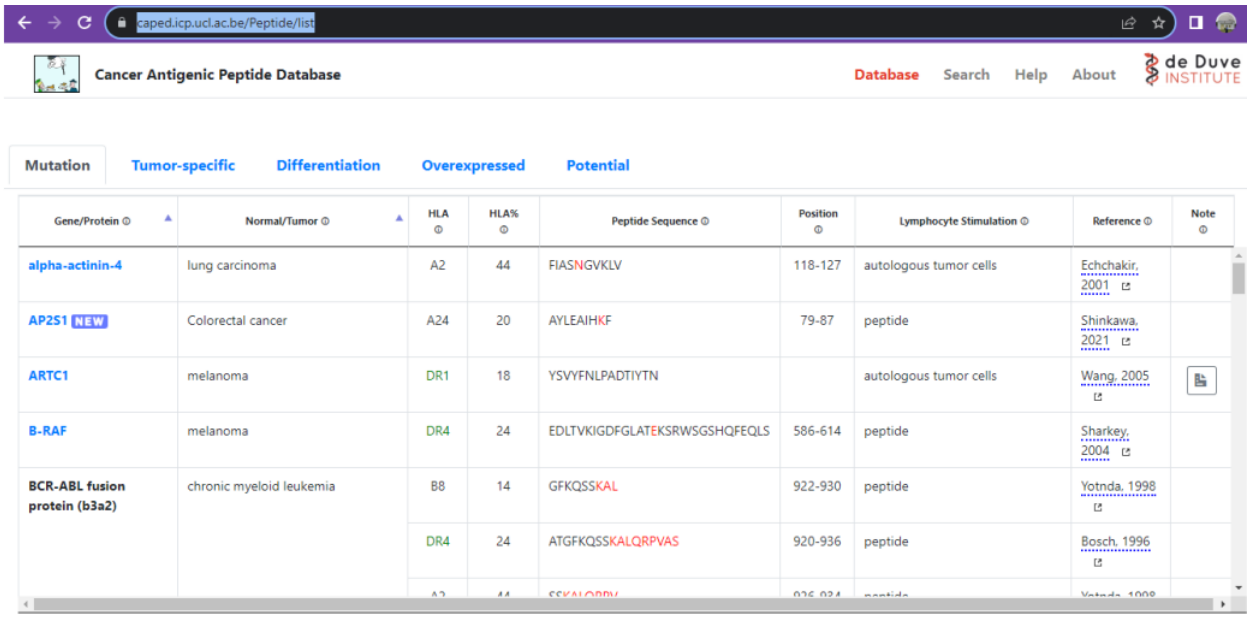


Figure 1. Research Workflow

2.1 Data Preprocessing

2.1.1 Data Collection

The dataset used in this study consists of cancer-related epitope sequences obtained from the Cancer Antigen Peptide Database (CAPED) [16]. Only validated peptide sequences are selected to ensure data quality. A total of 235 epitopes were collected, comprising 171 protein-derived epitope and 64 epitope mutations.



Gene/Protein	Normal/Tumor	HLA	HLA%	Peptide Sequence	Position	Lymphocyte Stimulation	Reference	Note
alpha-actinin-4	lung carcinoma	A2	44	FIASNGVKLV	118-127	autologous tumor cells	Echchakir, 2001	
AP2S1 NEW	Colorectal cancer	A24	20	AYLEAIHKF	79-87	peptide	Shinkawa, 2021	
ARTC1	melanoma	DR1	18	YSVYFNLPADTIYTN		autologous tumor cells	Wang, 2005	
B-RAF	melanoma	DR4	24	EDLTVKIGDFGLATEKSRWSGSHQFEQLS	586-614	peptide	Sharkey, 2004	
BCR-ABL fusion protein (b3a2)	chronic myeloid leukemia	B8	14	GFKQSSKAL	922-930	peptide	Yotnda, 1998	
		DR4	24	ATGFKQSSKALQRPVAS	920-936	peptide	Bosch, 1996	

Figure 2. Cancer Antigenic Peptide Database

Copyright © 2026 **Author**, Page 365
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

Figure 2 is the Cancer antigen database that is available online at https://caped.icp.ucl.ac.be/Peptide/list. Then, the result example of the data collection can be seen at Table 1. Epitope mutation in cancer refers to genetic alterations in tumor-derived antigens that change the regions recognized by the immune system, particularly by T cells [17]. These mutations often arise from genomic instability, a hallmark of cancer, leading to the formation of new epitopes that were not previously recognized as self.

While some new epitopes can stimulate strong anti-tumor immune responses, tumor cells may also accumulate mutations that alter or eliminate these epitopes, enabling immune evasion. This process contributes to tumor progression and resistance to treatment in diseases such as cancer. Recent studies highlight that epitope mutation and neoantigen evolution play critical roles in shaping tumor–immune interactions and influencing the success of immunotherapy [18].

Table 1. Epitopes Data Sample

Epitope Sequence	Protein/Gene
SLIAAAAFCLA	Cyclin-A1
RQKRILVNL	KK-LC-1
FLGPWAAS	LRPAP1
GLYDGMEHL	MAGE-A10
ALSVMGVYV	MAGE-A9
LYATVIHDI	SAGE

Table 2 is an example of a mutation in epitopes, the letters in red are the mutated epitope sequence.

Table 2. Mutation Epitope Data Sample

Epitope Sequence	Protein
SYLDSGIHF	beta-catenin
FPSDSWCYF	CASP-8
ACDPHSGHFV	CDK4
TMKQICKKEIRRLHQY	dek-can fusion protein
YVDFREYEYY	FLT3-ITD

2.1.2 Data Cleaning

The initial dataset consisted of 235 epitope samples. During preprocessing, one sample containing missing values (NaN) was removed, resulting in a final dataset of 234 samples. The final dataset comprised 170 tumor epitopes and 64 mutated tumor epitopes. This class distribution indicates a moderate imbalance between classes, which may influence model learning behavior and classification performance. In particular, machine learning models may become biased toward the majority class, potentially reducing their ability to correctly identify minority class instances. Therefore, the class distribution was taken into consideration during model evaluation by employing multiple performance metrics rather than relying solely on accuracy.

2.2 Sequence-derived Feature Analysis

Bioinformatics tools were employed to derive informative features from protein sequences, with Biogem specifically utilized for the analysis of amino acid physicochemical properties.

The physicochemical properties of each epitope were analyzed using the BioGEM platform as seen in Figure 3. The extracted attributes included molecular weight, theoretical isoelectric point (pI), instability index, aliphatic index, aromaticity, and grand average of hydropathicity (GRAVY), which were subsequently used as feature variables in this study [19]. These parameters are widely utilized to characterize protein and peptide properties, as they provide essential information regarding molecular stability, charge distribution, hydrophobicity, and structural behavior. Specifically, the instability index is commonly applied to predict protein stability, the aliphatic index reflects thermostability, and GRAVY indicates the hydrophobic or hydrophilic nature of a peptide [20]. Collectively, these physicochemical descriptors play an important role in understanding epitope characteristics and their potential biological functions, particularly in the context of immunogenicity and molecular interactions.

Following feature extraction, each epitope sequence was represented as a numerical feature vector composed of the calculated physicochemical descriptors. This transformation enabled the conversion of biological sequence information into a structured dataset suitable for machine learning analysis. Since classification algorithms cannot directly process amino acid sequences in textual form, the extracted features served as quantitative representations of the underlying biochemical characteristics of each epitope.

The resulting feature matrix was subsequently used as input for both Random Forest and Support Vector Machine models to distinguish between tumor and mutation epitopes. By utilizing physicochemical descriptors rather than raw sequence data, the classification process focused on identifying patterns associated with the biochemical properties of the epitopes that may contribute to their biological and immunological differences.

Copyright © 2026 **Author**, Page 366
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

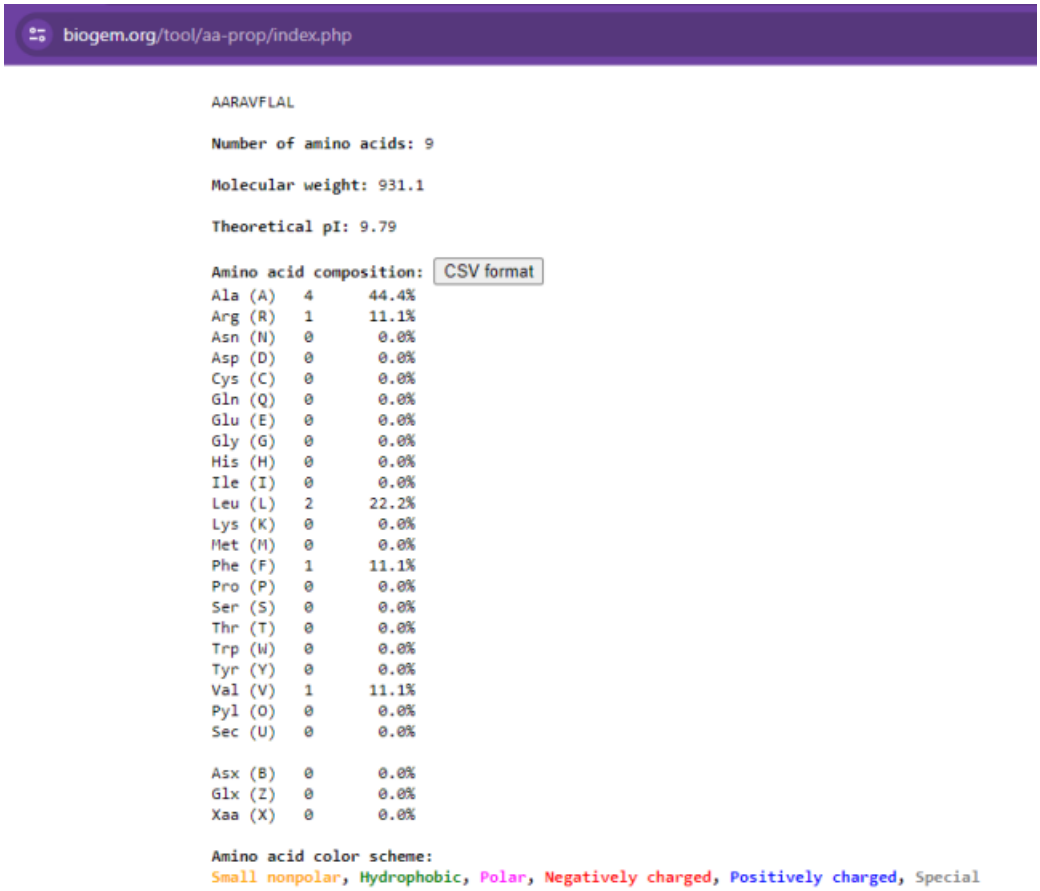


Figure 3. Biogem Platform

Table 3 presents several examples of epitope sequences and their corresponding physicochemical features extracted using the BioGEM platform. Each peptide sequence was represented by a set of numerical descriptors, including molecular weight, isoelectric point, instability index, aliphatic index, aromaticity, and GRAVY. In addition, the target column indicates the class label assigned to each epitope, where 0 represents tumor epitopes and 1 represents mutated tumor epitopes. These numerical features served as the input variables for the machine learning models, providing a structured representation of epitope characteristics for classification analysis.

Table 3. Epitope-Derived Feature

Peptide Sequence	Molecular Weight	PI	II	Aliphatic	Aromatic	Gravy	Target
AARAVFLAL	931,1	9,79	8.89	163.33	0.111	1.92	0
YLCSGSSYFV	1125.3	5.52	28.26	68	0.3	790	0
SLIAAAAFCLA	1050.3	5.24	26.6	151.82	0.091	2.327	0
FLDRFLSCM	1131.4	5.83	72.96	86.67	0.222	978	0
SLMEPPAVLLL	1182.5	4	61.62	177.27	0	1.418	0
VLPDVFIRC(V)	1160.4	5.8	13.18	165	0.1	1.66	0
FIASNGVKLV	1047.3	8.75	-17.4	146	0.1	1.27	1
AYLEAIHKF	1091.3	6.79	48.91	108.89	0.22	0.331	1
YSVYFNLPADTIYTN	1781.0	3.80	-1.37	78.00	0.27	-0.073	1
GFKQSSKAL	965.1	10.00	88.72	54.44	0.111	-0.544	1

2.3 Machine Learning Classification

This study applies two supervised learning algorithms, namely Random Forest and Support Vector Machine(SVM).

2.3.1 Random Forest (RF)

The classification model in this study is developed using the Random Forest algorithm, an ensemble learning method that constructs multiple decision trees and combines their predictions to improve classification performance. Each tree is trained using a bootstrap sample of the dataset, while a random subset of features is selected at each split to increase model diversity and reduce overfitting [21] [22]. The final prediction of the Random Forest model is determined majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_n(x)\} \quad (1)$$

Copyright © 2026 **Author**, Page 367
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

where $T_i(x)$represents the prediction of the i-th decision tree and $\hat{y}$denotes the final predicted class.

2.3.2 Support Vector Machine (SVM)

In addition to Random Forest, this study employs the Support Vector Machine (SVM) algorithm as a comparative classification model. SVM is a supervised learning method that aims to find the optimal hyperplane that separates data points from different classes with the maximum margin. Due to its effectiveness in handling high-dimensional data, SVM is widely used in bioinformatics and cancer-related classification tasks. For linearly separable data, the decision boundary is defined as:

$$w \cdot x + b = 0 \tag{2}$$

where $w$represents the weight vector and $b$is the bias term. The objective of SVM is to maximize the margin between classes while minimizing classification errors [23]. For non-linear data, SVM utilizes kernel functions to map input data into a higher-dimensional feature space. Commonly used kernels include linear, polynomial, and radial basis function (RBF). In this study, the RBF kernel is applied due to its ability to model complex biological patterns [24], [25]. Recent studies demonstrate that SVM continues to perform competitively in classification problems involving biological and medical datasets. For example, a study published in *Computers in Biology and Medicine* (2024) showed that SVM achieved high accuracy in cancer classification tasks when combined with feature selection techniques, highlighting its effectiveness in handling complex biological features [26].

2.4 Bioinformatics Implementation

To support the biological interpretation of the machine learning classification results, structural visualization of selected epitope regions was performed using available protein structural data from the Protein Data Bank (PDB) [27]. Pymol software was used as a representative model for illustrating the approximate localization of mutation-associated epitopes. The structural interpretation was intended to provide additional insight into the positional characteristics of mutation regions within the protein structure. This analysis was used solely for qualitative biological interpretation and was not intended as a comprehensive molecular docking or structural dynamics study. Visualization and structural inspection were conducted to identify whether mutation-associated epitopes were located within biologically relevant regions that may influence molecular interaction and immune recognition behavior.

3. RESULT AND DISCUSSION

3.1 Experimental Result

This study evaluated the performance of the Random Forest and Support Vector Machine (SVM) algorithms for classifying tumor and mutation data using different training and testing data split ratios. The experiments were conducted using three different data partition scenarios, namely 80:20, 75:25, and 70:30. The purpose of applying several data splitting configurations was to analyze the stability and effectiveness of each classification algorithm when trained with different amounts of data.

The performance evaluation was based on several classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. These metrics were selected to provide a comprehensive evaluation of the model performance, especially in identifying both majority and minority classes.

Table 4 presents the comparison results between the Random Forest and SVM algorithms under different data splitting scenarios. The experimental results demonstrate that the Random Forest algorithm consistently outperformed the SVM algorithm across all data splitting configurations. The superior performance of Random Forest may be attributed to its ensemble-learning mechanism, which aggregates the predictions of multiple decision trees and reduces model variance. This characteristic enables Random Forest to better capture non-linear relationships among physicochemical features while maintaining robustness against overfitting. In contrast, SVM performance is highly dependent on the selection of kernel parameters and may be more sensitive to overlapping feature distributions between classes.

Table 4. Comparison Between SVM and Random Forest in various Data Splitting

Training:Testing Data	Random Forest	SVM
80:20	83%	77%
75:25	80%	73%
70:30	76%	73%

The highest classification accuracy was obtained using the 80:20 training-testing split, where Random Forest achieved 83% accuracy, while SVM achieved 77% accuracy. Similarly, in the 75:25 split ratio, Random Forest obtained 80% accuracy compared to 73% for SVM. In the 70:30 split ratio, Random Forest achieved 76% accuracy, whereas SVM remained at 73%.

The results indicate that increasing the amount of training data positively affected model performance. The 80:20 split ratio provided the best results because the model had access to a larger amount of training data, enabling

Copyright © 2026 **Author**, Page 368
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

it to learn the patterns and characteristics of the dataset more effectively. In contrast, reducing the amount of training data in the 75:25 and 70:30 scenarios slightly decreased the classification accuracy, particularly for the Random Forest model.

3.2 Confusion Matrix Analysis of Random Forest and SVM

Due to the similarity of the classification patterns across different data split scenarios, only the confusion matrices corresponding to the best-performing models are presented in this paper. The remaining confusion matrices produced comparable results and therefore are not discussed in detail.

	precision	recall	f1-score	support
0	0.85	0.95	0.90	37
1	0.67	0.40	0.50	10
accuracy			0.83	47
macro avg	0.76	0.67	0.70	47
weighted avg	0.81	0.83	0.81	47

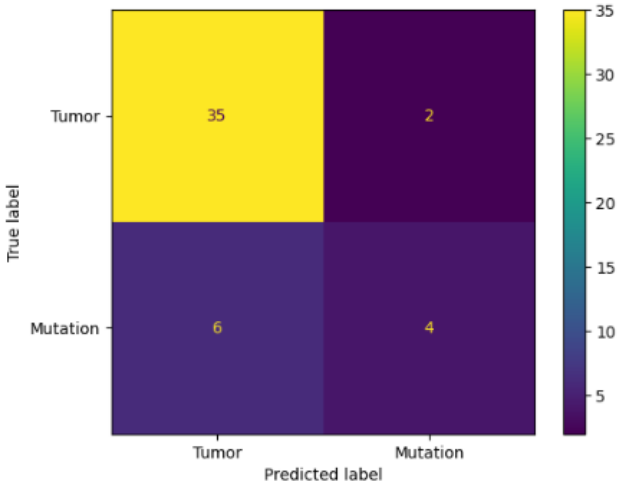


Figure 4. Confusion Matrix Random Forest with 80:20 data split

The confusion matrix for evaluating Random Forest model with 80% data training and 20% data testing illustrated in Figure 4. The model achieved an overall accuracy of 83%, indicating good classification performance on the testing dataset. The classification report further supports this observation. For the tumor class (Class 0), the model achieved a precision of 0.85, recall of 0.95, and F1-score of 0.90. The high recall value demonstrates that the Random Forest model was highly effective in detecting tumor samples with minimal false negatives. In contrast, the mutation class (Class 1) obtained a precision of 0.67, recall of 0.40, and F1-score of 0.50, indicating weaker performance in identifying mutation samples. The lower recall for the mutation class suggests that the model had difficulty recognizing mutation instances, likely due to class imbalance in the dataset, where tumor samples (37 instances) significantly outnumbered mutation samples (10 instances). As a result, the classifier tended to favor the majority class during prediction.

The model in Figure 5 was developed to classify samples into two categories: Tumor and Mutation with SVM. Overall, the model achieved an accuracy of 77% on the testing dataset, indicating that the proposed approach provides relatively reliable classification performance for both classes. The confusion matrix demonstrates that the model correctly classified 30 Tumor samples and 6 Mutation samples. However, there were several misclassifications, including 7 Tumor samples predicted as Mutation and 4 Mutation samples predicted as Tumor. These results suggest that the model performs better in recognizing the Tumor class compared to the Mutation class.

Based on the classification report, the Tumor class (Class 0) achieved a precision of 0.88, recall of 0.81, and F1-score of 0.85. The high precision value indicates that most samples predicted as Tumor were correctly identified, while the recall value shows that the majority of actual Tumor samples were successfully detected by the model. The strong F1-score further confirms balanced performance for this class.

In contrast, the Mutation class (Class 1) obtained a precision of 0.46, a recall of 0.60, and F1-score of 0.52. Although the recall value indicates that the model was able to identify more than half of the Mutation samples, the relatively low precision suggests the presence of false positive predictions. This imbalance may be caused by the smaller number of Mutation samples in the dataset, where only 10 samples were available compared to 37 Tumor samples. Consequently, the model tends to be biased toward the majority class.

The macro-average F1-score of 0.68 reflects moderate overall classification capability across both classes, whereas the weighted-average F1-score of 0.78 indicates that the dominant Tumor class strongly influences the overall performance metrics. These findings imply that the dataset imbalance significantly affects the classifier's ability to generalize equally across classes.

Copyright © 2026 **Author**, Page 369
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

	precision	recall	f1-score	support
0	0.88	0.81	0.85	37
1	0.46	0.60	0.52	10
accuracy			0.77	47
macro avg	0.67	0.71	0.68	47
weighted avg	0.79	0.77	0.78	47

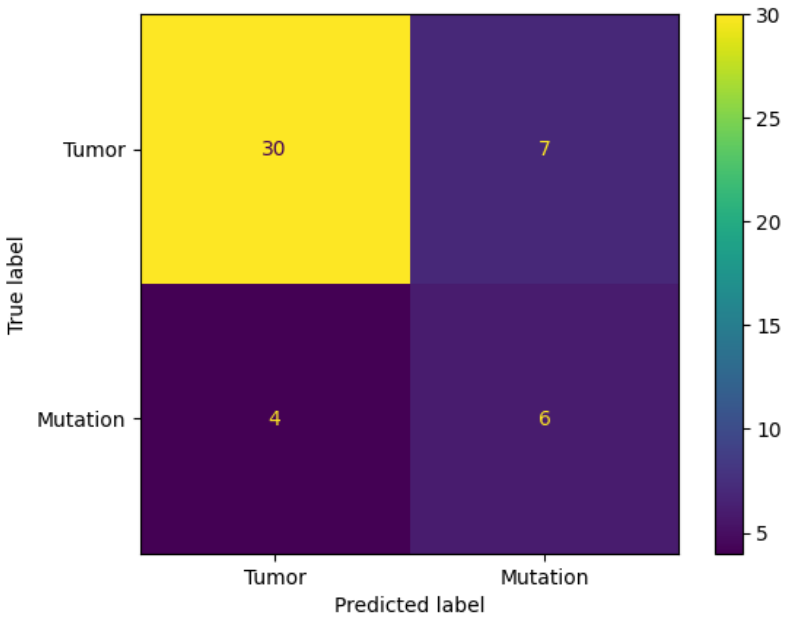


Figure 5. Confusion Matrix SVM with 80:20 data split

3.3 Bioinformatics Implementations

Figure 6 illustrates the structural localization of the predicted epitope region within the ACTN4 protein structure (PDB ID: 6O31) [28]. The identified epitope sequence, FIASKGVKLV, is located at amino acid positions 118–127. Within this epitope region, a mutation was observed at position 122, where the original amino acid lysine (K) was substituted by asparagine (N). The epitope region is highlighted in yellow, while the mutation-associated residue is indicated in orange.

The findings of this study also suggest that mutation-associated epitope classification cannot rely solely on generalized sequence characteristics. In biological systems, immune recognition processes are highly dependent on specific amino acid arrangements and positional interactions within peptide sequences. Consequently, even minor localized substitutions may produce biologically meaningful effects despite minimal changes in overall sequence composition.

In the ACTN4 epitope analyzed in this study, the substitution of lysine (K) with asparagine (N) occurred within a confined sequence region that may contribute to peptide recognition behavior. Since lysine possesses a positively charged side chain whereas asparagine is polar but electrically neutral, the substitution may alter local electrostatic interactions and potentially influence peptide–receptor binding behavior.

These observations further indicate that mutation-associated sequence variations may contain biologically important contextual information that is difficult to represent using conventional averaged physicochemical descriptors. Therefore, future machine learning approaches may benefit from incorporating sequence-context encoding, positional embeddings, or structure-aware feature extraction techniques to improve mutation epitope discrimination performance.

The structural localization observed in Figure 6 also provides a possible explanation for the classification results obtained in this study. Although the mutation occurs within an immunologically relevant epitope region, its impact is confined to a single amino acid position and does not substantially alter the overall physicochemical profile of the peptide. Consequently, the extracted features, which primarily describe global sequence characteristics, may remain relatively similar between tumor and mutated epitopes. This similarity can reduce the discriminative power of conventional machine learning algorithms and contribute to the observed false-negative predictions for the mutation class. Therefore, the structural findings support the hypothesis that local sequence alterations may be insufficiently represented by global descriptors alone, emphasizing the need for more biologically informative feature representations in future mutation epitope classification studies.

Copyright © 2026 **Author**, Page 370
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

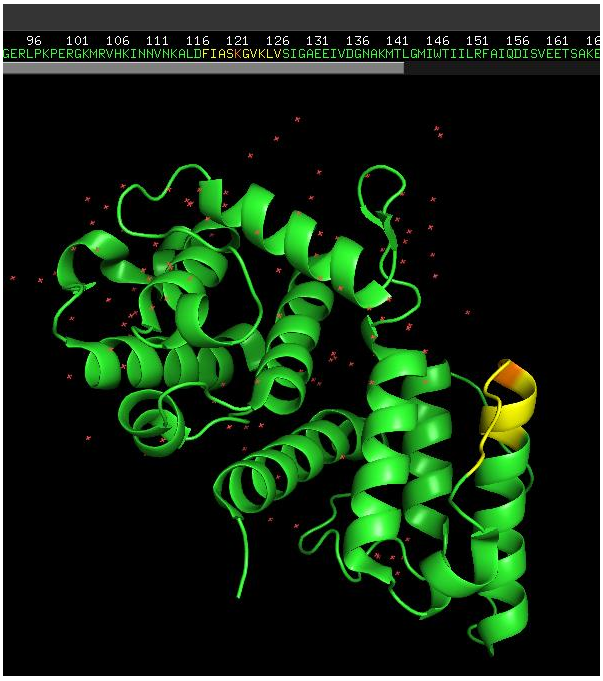


Figure 6. Molecular modelling

The ACTN4 protein comprises of a long amino acid sequence containing multiple structural and functional regions. In this study, the identified mutation-associated epitope FIASK(N)GVKLV represents only a small localized segment within the overall ACTN4 sequence as shown in Figure 7.

```
>NP_001308962.1 ACTN4 [organism=Homo sapiens] [GeneID=81] [isoform=2]
MVDYHAANQSYQYGPSSAGNGAGGGGSMGDYMAQEDDWDRDLLLDPAWEKQQRKTFTAWCNSHLR
KAGTQIENIDEDFRDGLKLMLLLEVISGERLPKPERGKMRVHKINNVNKALDFIASKGVKLVSIG
AEEIVDGNAKMTLGMIWTIILRFAIQDISVEETSAKEGLLLWCQRKTAPYKNVNVQNFHISWKDG
LAFNALIHRHRPELIEYDKLRKDDPVTNLNNAFEVAEKYLDIPKMLDAEDIVGTLRPDEKAIMTY
VSCFYHAFSGAQKAETAANRICKVLAVNQENEHLMEDYEKLASDLLEWIRRTIPWLEDRVPQKTI
QEMQQKLEDFRDYRRVHKPPKVQEKCQLEINFNTLQTKLRLSNRPAFMPSEGKMVSDINNGWQHL
EQAEKGYEEWLLNEIRRLERLDHLAEKFRQKASIHEAWTDGKEAMLKHRDYETATLSDIKALIRK
HEAFESDLAAHQDRVEQIAAIAQELNELDYYDSHNVNTRCQKICDQWDALGSLTHSRREALEKTE
KQLEAIDQLHLEYAKRAAPFNNWMESAMEDLQDMFIVHTIEEIEGLISAHDQFKSTLPDADRERE
AILAIHKEAQRIAESNHIKLSGSNPYTTVTPQIINSKWEKVQQLVPKRDHALLEEQSKQQSNEHL
RRQFASQANVVGPWIQTKMEEIGRISIEMNGTLEDQLSHLKQYERSIVDYKPNLDLLEQQHQLIQ
EALIFDNKHTNYTMEHIRVGWEQLLTTIARTINEVENQILTRDAKGISQEQMQEFRASFNHFDKK
QTGSMDSDDFRALLISTGYSLGEAEFNRIMSLVDPNHSGLVTFQAFIDFMSRETTDTDTADQVIA
SFKVLAGDKNFITAEELRRELPPDQAEYCIARMAPYQGPDAVPGALDYKSFSTALYGESDL
```

Figure 7. ACTN4 Amino Acids Sequences

The substitution from lysine (K) to asparagine (N) at residue position 122 may alter the local physicochemical environment of the epitope region. Lysine is a positively charged amino acid, whereas asparagine is polar but electrically neutral. Such changes may influence local peptide behavior, molecular interaction properties, and immune recognition patterns. In biological systems, even single amino acid substitutions can affect peptide conformation and antigen presentation mechanisms, potentially contributing to differences in epitope immunogenicity despite relatively small changes in overall sequence composition.

The structural visualization demonstrates that the mutation occurs within a localized region of the epitope rather than across the entire protein structure. This observation supports the notion that mutation-associated alterations are highly position-specific and may not substantially affect the global physicochemical characteristics of the protein. Consequently, feature extraction approaches relying primarily on averaged physicochemical descriptors may fail to capture subtle but biologically important sequence variations.

This positional characteristic may explain the lower classification performance observed for mutation epitopes in the machine learning experiments. Since the mutation effect occurs locally, localized sequence variations may not have been fully represented during model training. Furthermore, the localization of mutation residues within biologically relevant epitope regions suggests that structural and positional information may contribute to immune recognition behavior.

Therefore, incorporating sequence-position or structure-aware feature representations may improve mutation epitope classification performance in future studies. Although the structural visualization presented in this study is qualitative in nature, it provides additional biological interpretation supporting the machine learning classification findings.

Copyright © 2026 **Author**, Page 371
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

3.4 Discussion and limitations of the study

Despite the relatively good overall performance, further analysis reveals that the model exhibits limitations in correctly identifying mutated tumor epitopes. This is reflected in the relatively high number of false negatives, where mutated epitopes are incorrectly classified as non-mutated tumor epitopes. This issue is critical, as accurate identification of mutated epitopes is essential in cancer immunotherapy applications.

One possible explanation for this limitation lies in the nature of the features used in this study. The model relies on sequence-derived physicochemical properties, such as molecular weight, isoelectric point, aliphatic index, aromaticity, and hydrophobicity. These features represent the global characteristics of peptide sequences rather than local variations. As a result, point mutations occurring at specific amino acid positions may not significantly alter the overall feature values.

This observation is supported by the comparison of feature values between tumor and mutated tumor samples, where several mutated epitopes exhibit similar physicochemical properties to non-mutated epitopes. For instance, attributes such as aromaticity and aliphatic index show overlapping value ranges across both classes. Consequently, the model struggles to distinguish between the two classes, leading to misclassification.

In addition, the distribution of instability index values in mutated epitopes appears to be more variable, including both positive and negative values, which may increase intra-class variability. This further complicates the classification task, as the model must learn more complex patterns within the mutated class.

Another contributing factor is the potential imbalance in the dataset, where the number of mutated tumor samples may be lower than non-mutated samples. Imbalanced data can bias the model toward the majority class, resulting in higher false negative rates [29]. This suggests that future work should consider applying data balancing techniques, such as oversampling with SMOTE to improve model sensitivity toward mutated epitopes [25].

Furthermore, the findings highlight an important limitation of using only global sequence-derived features for mutation classification. Since mutations are inherently local changes in amino acid sequences, feature representations that capture positional or sequence-specific information may be more effective. Techniques such as amino acid composition, sequence encoding, or deep learning-based embeddings could potentially improve classification performance. Future improvements are still needed.

4. CONCLUSION

This study successfully applied machine learning techniques, namely Random Forest and Support Vector Machine (SVM), for the classification of cancer-related epitopes. The experimental results demonstrate that the Random Forest model outperforms SVM, achieving the highest accuracy of 83% using an 80:20 train-test split. Based on the confusion matrix analysis, the Random Forest model shows strong performance in correctly classifying tumor (non-mutated) epitopes, with a high number of true positive predictions. However, the model exhibits relatively lower performance in identifying mutated tumor epitopes, as indicated by the higher number of false negatives. This suggests that distinguishing mutated epitopes remains a challenging task, likely due to subtle differences in feature representation and potential class imbalance within the dataset. The study also reveals that model performance is influenced by the proportion of training data. A decrease in training data size leads to reduced accuracy, highlighting the importance of sufficient data for effective model learning and generalization. This research contributes to computational immunology by demonstrating the applicability of sequence-derived physicochemical features for mutation epitope classification and by providing a comparative evaluation of Random Forest and SVM under multiple data partition scenarios. Furthermore, the integration of machine learning results with structural bioinformatics interpretation offers additional biological insights into mutation-associated epitopes**.** Overall, the findings confirm that ensemble learning methods such as Random Forest are well-suited for handling complex biological data and can provide reliable performance in epitope prediction tasks. For future work, improvements can be made by incorporating advanced feature engineering techniques, applying data balancing methods, and exploring hybrid or ensemble approaches to enhance the detection of mutated epitopes. Additionally, increasing the dataset size and diversity may further improve model robustness and prediction accuracy.

REFERENCES

[1] Badan Pusat Statistik, *Statistik Indonesia*, vol. 53. 2025. [Online]. Available: https://www.bps.go.id/id/publication/2025/02/28/8cfe1a589ad3693396d3db9f/statistik-indonesia-2025.html

[2] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: https://doi.org/10.3322/caac.21834.

[3] M. J. Kamali, M. Salehi, and M. K. Fath, "Advancing personalized immunotherapy for melanoma: Integrating immunoinformatics in multi-epitope vaccine development, neoantigen identification via NGS, and immune simulation evaluation," *Comput. Biol. Med.*, vol. 188, p. 109885, Apr. 2025, doi: 10.1016/j.compbiomed.2025.109885.

[4] W. Chronister, A. Sette, and B. Peters, "Epitope-Specific T Cell Receptor Data and Tools in the Immune Epitope Database," in *T-Cell Repertoire Characterization*, H. Huang and M. M. Davis, Eds., New York, NY: Springer US, 2022, pp. 267–280. doi: 10.1007/978-1-0716-2712-9_13.

[5] G. Dolton *et al.*, "Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy," *Cell*, vol. 186, no. 16, pp. 3333-3349.e27, Aug. 2023, doi: 10.1016/j.cell.2023.06.020.

Copyright © 2026 **Author**, Page 372
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 363−373
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.1002

[6] H. Lin *et al.*, "Identification of Tumor Antigens and Immune Subtypes of Glioblastoma for mRNA Vaccine Development," *Front. Immunol.*, vol. 13, 2022, doi: 10.3389/fimmu.2022.773264.
[7] S. Grewal, N. Hegde, and S. K. Yanow, "Integrating machine learning to advance epitope mapping," *Front. Immunol.*, vol. 15, 2024, doi: 10.3389/fimmu.2024.1463931.
[8] F. Villanueva-Flores, J. I. Sanchez-Villamil, and I. Garcia-Atutxa, "AI-driven epitope prediction: a systematic review, comparative analysis, and practical guide for vaccine development," *Npj Vaccines*, vol. 10, no. 1, p. 207, Aug. 2025, doi: 10.1038/s41541-025-01258-y.
[9] B. Dou *et al.*, "Machine Learning Methods for Small Data Challenges in Molecular Science," *Chem. Rev.*, vol. 123, no. 13, pp. 8736–8780, Jul. 2023, doi: 10.1021/acs.chemrev.3c00189.
[10] P. Gautam and P. Mitra, "ConfPred: ML-Based Conformational B-cell Epitope Prediction Using Novel Features," in *Proceedings of the 12th International Conference on Bioinformatics Research and Applications*, in ICBRA '25. New York, NY, USA: Association for Computing Machinery, Dec. 2025, pp. 103–110. doi: 10.1145/3774976.3774994.
[11] W. Alghamdi, M. Attique, E. Alzahrani, M. Z. Ullah, and Y. D. Khan, "LBCEPred: a machine learning model to predict linear B-cell epitopes," *Brief. Bioinform.*, vol. 23, no. 3, p. bbac035, May 2022, doi: 10.1093/bib/bbac035.
[12] T.-O. Tran and N. Q. K. Le, "Sa-TTCA: An SVM-based approach for tumor T-cell antigen classification using features extracted from biological sequencing and natural language processing," *Comput. Biol. Med.*, vol. 174, p. 108408, May 2024, doi: 10.1016/j.compbiomed.2024.108408.
[13] N. Flechas Manrique *et al.*, "epiGPTope: A Machine Learning-Based Epitope Generator and Classifier," *ACS Synth. Biol.*, vol. 15, no. 2, pp. 631–642, Feb. 2026, doi: 10.1021/acssynbio.5c00693.
[14] R.-S. Hu, K. Gu, M. Ehsan, S. H. A. Raza, and C.-R. Wang, "Transformer-based deep learning enables improved B-cell epitope prediction in parasitic pathogens: A proof-of-concept study on Fasciola hepatica," *PLoS Negl. Trop. Dis.*, vol. 19, no. 4, p. e0012985, Apr. 2025, doi: 10.1371/journal.pntd.0012985.
[15] A. Albutti, "An Integrated Approach to Develop a Potent Vaccine Candidate Construct Against Prostate Cancer by Utilizing Machine Learning and Bioinformatics," *Cancer Rep.*, vol. 7, no. 12, p. e70079, 2024, doi: 10.1002/cnr2.70079.
[16] N. Biswas, S. Chakrabarti, V. Padul, L. D. Jones, and S. Ashili, "Designing neoantigen cancer vaccines, trials, and outcomes," *Front. Immunol.*, vol. 14, 2023, doi: 10.3389/fimmu.2023.1105420.
[17] S. Feola, J. Chiaro, and V. Cerullo, "Integrating immunopeptidome analysis for the design and development of cancer vaccines," *Semin. Immunol.*, vol. 67, p. 101750, May 2023, doi: 10.1016/j.smim.2023.101750.
[18] M. Łuksza *et al.*, "A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy," *Nature*, vol. 551, no. 7681, pp. 517–520, Nov. 2017, doi: 10.1038/nature24473.
[19] P. Karnati *et al.*, "Performance of Novel Antimicrobial Protein Bg_9562 and In Silico Predictions on Its Properties with Reference to Its Antimicrobial Efficiency against Rhizoctonia solani," *Antibiotics*, vol. 11, no. 3, p. 363, Mar. 2022, doi: 10.3390/antibiotics11030363.
[20] S. Kumar, V. K. Duggineni, V. Singhania, S. P. Misra, and P. A. Deshpande, "Unravelling and Quantifying the Biophysical–Biochemical Descriptors Governing Protein Thermostability by Machine Learning," *Adv Theory Simul.*, vol. 6, no. 3, 2023, doi: 10.1002/adts.202200703.
[21] H. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylon. J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
[22] A. Y. Mahmoud, "Novel efficient feature selection: Classification of medical and immunotherapy treatments utilising Random Forest and Decision Trees," *Intell.-Based Med.*, vol. 10, p. 100151, Jan. 2024, doi: 10.1016/j.ibmed.2024.100151.
[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[24] L. Revathi and R. Murugesh, "A review of support vector machine in cancer prediction on genomic data," *Int. J. Bioinforma. Res. Appl.*, vol. 20, no. 2, pp. 161–180, Jan. 2024, doi: 10.1504/IJBRA.2024.138709.
[25] G. O. Anyanwu, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "RBF-SVM kernel-based model for detecting DDoS attacks in SDN integrated vehicular network," *Ad Hoc Netw.*, vol. 140, p. 103026, Mar. 2023, doi: 10.1016/j.adhoc.2022.103026.
[26] T. Nadira and Z. Rustam, "Classification of cancer data using support vector machines with features selection method based on global artificial bee colony," presented at the AIP Conference Proceedings, Oct. 2018, p. 020205. doi: 10.1063/1.5064202.
[27] S. K. Burley *et al.*, "RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D488–D508, Jan. 2023, doi: 10.1093/nar/gkac1077.
[28] D. Feng *et al.*, "Phosphorylation of ACTN4 Leads to Podocyte Vulnerability and Proteinuric Glomerulosclerosis," *J. Am. Soc. Nephrol.*, vol. 31, no. 7, p. 1479, Jul. 2020, doi: 10.1681/ASN.2019101032.
[29] P. Rana, A. Sowmya, E. Meijering, and Y. Song, "Imbalanced classification for protein subcellular localization with multilabel oversampling," *Bioinformatics*, vol. 39, no. 1, p. btac841, Jan. 2023, doi: 10.1093/bioinformatics/btac841.
[30] E. Jafarigol, T. Trafalis, and N. Mohammadi, "A Review of Machine Learning Techniques in Imbalanced Data and Future Trends," Sep. 07, 2025, *arXiv*: arXiv:2310.07917. doi: 10.48550/arXiv.2310.07917.

This Journal is licensed under a Creative Commons Attribution 4.0 International License