8-23-2022

C10Pred: A first machine learning based tool to predict C10 family cysteine peptidases using sequence-derived features

Adeel Malik

Nitin Mahajan

Tanveer Ali Dar

Chang-Bae Kim

International Journal of
Molecular Sciences

MDPI

Article

C10Pred: A First Machine Learning Based Tool to Predict C10 Family Cysteine Peptidases Using Sequence-Derived Features

Adeel Malik [1,*], Nitin Mahajan [2,†], Tanveer Ali Dar [3] and Chang-Bae Kim [4,*]

[1] Institute of Intelligence Informatics Technology, Sangmyung University, Seoul 03016, Korea
[2] Department of Pediatrics, Washington University in St. Louis, St. Louis, MO 63110, USA
[3] Department of Clinical Biochemistry, University of Kashmir, Srinagar 190006, India
[4] Department of Biotechnology, Sangmyung University, Seoul 03016, Korea
* Correspondence: adeel@procarb.org (A.M.); evodevo@smu.ac.kr (C.-B.K.)
† Present Address: Wugen, St. Louis, MO 63110, USA.

Abstract: *Streptococcus pyogenes*, or group A *Streptococcus* (GAS), a gram-positive bacterium, is implicated in a wide range of clinical manifestations and life-threatening diseases. One of the key virulence factors of GAS is streptopain, a C10 family cysteine peptidase. Since its discovery, various homologs of streptopain have been reported from other bacterial species. With the increased affordability of sequencing, a significant increase in the number of potential C10 family-like sequences in the public databases is anticipated, posing a challenge in classifying such sequences. Sequence-similarity-based tools are the methods of choice to identify such streptopain-like sequences. However, these methods depend on some level of sequence similarity between the existing C10 family and the target sequences. Therefore, in this work, we propose a novel predictor, C10Pred, for the prediction of C10 peptidases using sequence-derived optimal features. C10Pred is a support vector machine (SVM) based model which is efficient in predicting C10 enzymes with an overall accuracy of 92.7% and Matthews' correlation coefficient (MCC) value of 0.855 when tested on an independent dataset. We anticipate that C10Pred will serve as a handy tool to classify novel streptopain-like proteins belonging to the C10 family and offer essential information.

Keywords: C10 family; cysteine peptidase; streptopain; machine learning; support vector machine; feature selection; Boruta

Citation: Malik, A.; Mahajan, N.; Dar, T.A.; Kim, C.-B. C10Pred: A First Machine Learning Based Tool to Predict C10 Family Cysteine Peptidases Using Sequence-Derived Features. *Int. J. Mol. Sci.* **2022**, *23*, 9518. https://doi.org/10.3390/ijms23179518

Academic Editor: Efstratios Stratikos

Received: 14 July 2022
Accepted: 20 August 2022
Published: 23 August 2022

1. Introduction

Streptococcus pyogenes, or group A *Streptococcus* (GAS), a gram-positive bacterium, is a reason for causing a wide range of clinical manifestations, from superficial infections to severe, life-threatening diseases [1,2]. Globally, GAS pharyngitis accounts for more than 600 million cases annually [3]. Another superficial infection frequently caused by GAS is impetigo, commonly found in tropical, resource-poor settings [4]. It has been estimated that more than 162 million children globally suffer from impetigo at any time [5]. On the other hand, GAS causes severe and life-threatening infections, such as bacteremia, meningitis, necrotizing fasciitis (NF), sepsis, and streptococcal toxic shock syndrome (STSS) [6,7]. Furthermore, GAS infection can also develop immune-mediated sequelae like acute rheumatic fever (ARF) and acute post-streptococcal glomerulonephritis (APSGN) [8]. ARF can lead to rheumatic heart disease (RHD), a significant cause of mortality and morbidity resulting from the severe condition associated with cardiac failure, stroke, and early death [8]. APSGN can contribute to chronic renal failure [9,10]. Although now rare in developed world countries, ARF and APSGN still maintain a significant presence in economically disadvantaged populations accounted for by poor hygiene and limited resources [11,12]. However, relatively rare invasive infections are often associated with high mortality and morbidity [3]. The global incidence of invasive GAS diseases is reportedly around 6 cases per 100,000 people per year [13]. The highest incidence of invasive GAS diseases is seen in

the elderly, followed by young children, particularly those under one year of age [14]. GAS infections are significantly associated with high morbidity and mortality rates worldwide. GAS diseases were estimated to cause more than 500,000 deaths in 2005 [3], and RHD alone was estimated to account for 320,000 deaths in 2015 [15].

Streptopain, also known as streptococcal pyrogenic exotoxin B (SpeB), streptococcus peptidase A (SPP), and streptococcal cysteine protease (SCP), is one of the major virulence factors for GAS [16–18]. According to MEROPS [19], an online peptidase database, streptopain is classified as a C10 family (also known as the streptopain family) of cysteine peptidases. There are 16 clans of cysteine peptidases, including some unclassified, of which four comprises proteases with diverse catalytic types [19]. Each of these clans exhibits a distinct fold and is further divided into separate families [20]. Streptopain belongs to Clan CA of cysteine proteases and shares the clan with the first-ever discovered cysteine protease, i.e., papain (isolated from *Carica papaya*). Cysteine proteases use the reactive site cysteine as the catalytic nucleophile and the histidine to perform peptide bond hydrolysis. Streptopain shares limited similarity with papain, and the catalytic cysteine and histidine residues in streptopain (C47and H195) have similar order to papain (C25 and H159), including some identical neighboring residues [19,20]. Sequence analysis of all intermediates and final product demonstrates that streptopain, similar to papain, prefers substrates with hydrophobic residues [21]. However, streptopain lacks the presence of Asn or Asp residue equivalent to the Asn in the papain family, which forms the catalytic triad in cysteine peptidases. Also, in contrast to other papain family members, streptopain has significant insertions and deletions outside its conserved core. The pro-domain of streptopain has a fold that is unique among the other proteases [22].

Streptopain is present in all the isolates of *S. pyogenes*, and therefore is a predominant extracellular protein that accounts for approximately 95% of total secreted proteins [23]. Streptopain is extracellularly released from GAS to the culture medium in a zymogen form, i.e., proSPEB [24,25]. Zymogen (proSPEB) is converted to the mature mSPEB either by autoproteolysis or exogenous proteases. Structurally, the conformation of the C-terminal loop and the orientation of the catalytic H195 residue plays an important role in activating proSPEB to mSPEB [21]. NMR analysis demonstrates that the C-terminal loop of streptopain is flexible, controls the substrate binding, and therefore has diverse substrate specificity [22,26]. Streptopain has diverse substrate specificity in processing streptococcal proteins and host proteins. This diverse substrate specificity leads to its different biological effects [27]. For example, streptopain degrades extracellular matrix (ECM) proteins fibronectin and vitronectin, which help in bacterial attachment to host cells [28]. It also cleaves and activates matrix metalloproteases (MMPs) and therefore helps in extracellular matrix degradation, which eventually leads to increased bacterial dissemination [29,30]. In addition to host proteins, streptopain releases streptococcal surface proteins like M-protein, protein F1, protein H, Sda1, Fba, and superantigen1 [27]. Besides protease activity, streptopain also exhibits transferase and esterase activities. A variant of streptopain with an Arg-Gly-Asp motif that binds integrins $\alpha v\beta 3$ and $\alpha IIb\beta 3$ has been reported in the M1 serotype isolates [31].

Although much of the work on this C10 cysteine peptidase has been reported from Streptococcal strains, studies have also identified SpeB homologs from other bacterial species. Among them is interpain A (InpA), which was identified from an oral anaerobe *Prevotella intermedia*. InpA plays an essential role in the oxidation and breakdown of hemoglobin and the subsequent release of haem [32]. Similarly, two genes encoding periodontain (PdnA) and thiol protease/hemagglutinin (PrtT) from *Porphyromonas gingivalis* share significant homology to SpeB [33]. Streptopain homologs have also been discovered from bacterial species that inhabit organisms other than humans. This includes bacteria that are pathogenic to marine aquaculture. For example, the *FcpB* gene in *Flavobacterium psychrophilum*, a Gram-negative fish pathogen, encodes a 394 amino-acid protein fcpB [34]. This protein shares significant homology with cysteine peptidases such as streptopains and other C10 family members from different bacterial species, including *Flavobacterium*

branchiophilum, *Dyadobacter fermentans*, *Bacteroides intestinalis*, and *Spirosoma linguale* [34]. Similarly, a gene cluster *MARIT_2328* in *Tenacibaculum maritimum* encodes a multi-domain protein from the C10 family peptidase, which is significantly similar to SpeB, and likely plays a role in colonization and invasion [35]. The genomic overview of the peptidases of anaerobic Gram-negative bacteria *Prevotella* and *Paraprevotella* species which inhabit the oral cavity, GI tract, and urinary tract of animals and humans, provided a comprehensive analysis of various peptidases [36,37]. Genomic sequencing of *Prevotella* and *Paraprevotella* species demonstrated the presence of a total of 78 distinct peptidase families. This analysis shows that C10 family peptidases were among the most abundant [38].

Since its discovery, several streptopain homologs have been identified, many of which remain uncharacterized. Additionally, the large-scale bacterial genome sequencing projects may have led to a rapid increase in the number of potential C10 family-like sequences in the public databases, posing a challenge to annotating such sequences. Moreover, experimental identification and characterization of streptopain and its homologs is costly and time-consuming. Therefore, novel computational methods are required that provide robust techniques for correctly identifying C10 family cysteine peptidases from their primary amino acid sequences. At present, methods based on sequence-similarity such as BLAST [39] and HMMER [40] are the only approaches that are available to identify streptopain-like sequences. However, one of the main drawbacks of such techniques is that they are meaningful only if there exists some level of sequence similarity between the existing C10 family and the target sequences. Consequently, these methods fail to discover novel sequences comprising streptopain-like domains. Hence, machine learning (ML)-based approaches offer encouraging alternatives to develop novel predictors for such classification problems.

With this background, we propose the first ML-based tool, C10Pred, which can predict C10 family proteins from their primary sequences. The predictor incorporates optimal features from different encodings (hybrid features) for better performance. We expect that C10Pred will be a competent tool for identifying the C10 family or streptopain-like sequences, which will help investigate their functional roles in many bacterial diseases. The C10Pred web server is freely available at https://procarb.org/c10pred/ (accessed on 16 August 2022).

2. Results

2.1. Overview of the Dataset

The positive dataset comprises a non-redundant set of 336 C10 family peptidases belonging to the PFAM family "Peptidase_C10", whereas the negative data consists of 350 sequences from other cysteine peptidase families within the MEROPS [19] database (Table 1). This dataset represents sequences from a wide range of taxonomic groups and corresponds to 82 and 283 unique bacterial taxonomic groups in the positive and negative datasets, respectively (Figure 1 and Table S1). The functional annotation of the dataset was carried out with an eggNOG mapper [41] and suggested that more than 80% of sequences in the positive dataset belong to an unknown functional category (S = 43.54%) or did not have any hits in the eggNOG database. In contrast, the annotation was much better for the negative dataset. Some of the well-annotated COG categories include posttranslational modification, protein turnover, chaperones (O = 18.85%), cell wall/membrane/envelope biogenesis (M = 12%), and amino acid transport and metabolism (E = 8.57%) (Figure S1). These data suggest that only limited information is available on the functional roles of C10 family peptidases and their homologs.

Table 1. A statistical summary of the training and independent datasets. * = Non-C10 family cysteine proteases; ** = All bacterial sequences except C10 family or streptopain proteins.

Class	Datasets			
	Training Set	Independent Validation Sets		
		VS1	VS2	VS3
Positive (C10 family cysteine proteases)	269	67	82	82
Negative	280	70	200 *	349 **

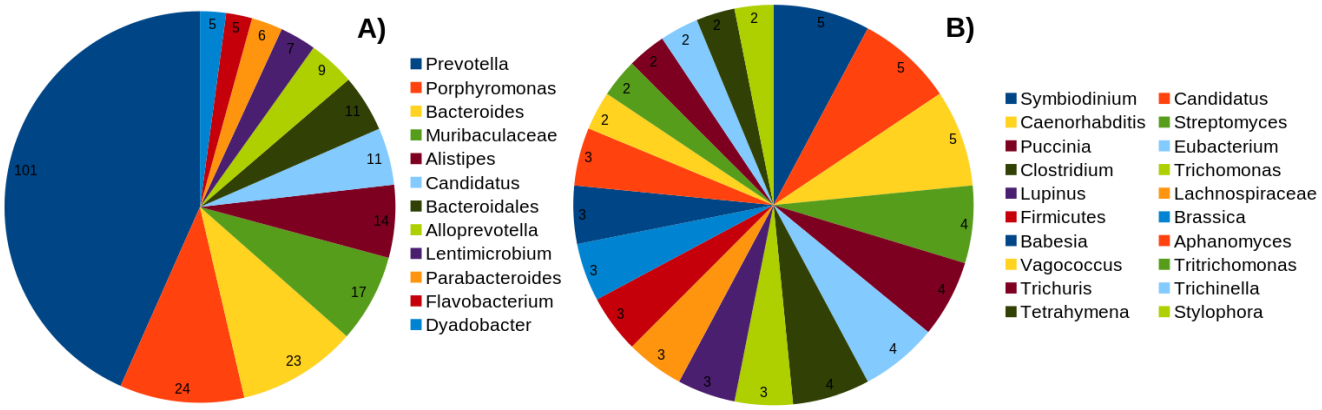

Figure 1. Number of most abundant taxonomic groups in the (**A**) positive, and (**B**) negative datasets. Complete list is provided in Table S1.

2.2. The Overall Framework of the Proposed Predictor

Figure 2 summarizes the overall framework for the C10Pred, which essentially consists of four main steps: (i) construction of training and independent datasets; (ii) encoding of various sequence-derived features (e.g., AAC, AutoC, CTD, CTriad, DPC, QSO, SOCN, and Hybrid); (iii) feature selection using Boruta algorithm, and (iv) selection of the final model exhibiting best performance in terms of MCC. Accordingly, the corresponding feature set with the best performance was considered to be the optimal set.

2.3. Amino Acid Composition in C10 and Non-C10 Sequences

To determine the presence of any compositional differences between C10 and non-C10 peptidases, we compared the AAC of the positive and negative datasets. The AAC of both these datasets is shown in Figure 3, which shows the higher frequency of hydrophobic amino acids like tryptophan (W), methionine (M), glycine (G), Isoleucine (I), and uncharged polar amino acids, including asparagine (N), threonine (T), and tyrosine (Y) in the C10 protein sequences (Wilcox test; $p < 0.05$) as compared to non-C10 peptidases. Interestingly, as compared to C10 peptidases with only aspartic acid as the dominant polar residue, the non-C10 were dominant in most of the charged polar residues, including arginine (R), glutamate (E), and histidine (H) residues. The most important characteristic feature for identification purposes could be the lesser frequency of charged polar residues and dominance of hydrophobic amino acids, particularly tryptophan, the rarest of the amino acids. The dominance of hydrophobic amino acids with a lesser frequency of charged residues in turn signifies the increased stability of C10 peptidases compared to non-C10 peptidases. The peculiar compositional differences between the peptidases in turn infer that our model could use the presence of specific amino acids as a suitable strategy to categorize C10 peptidases from non-C10 peptidases.

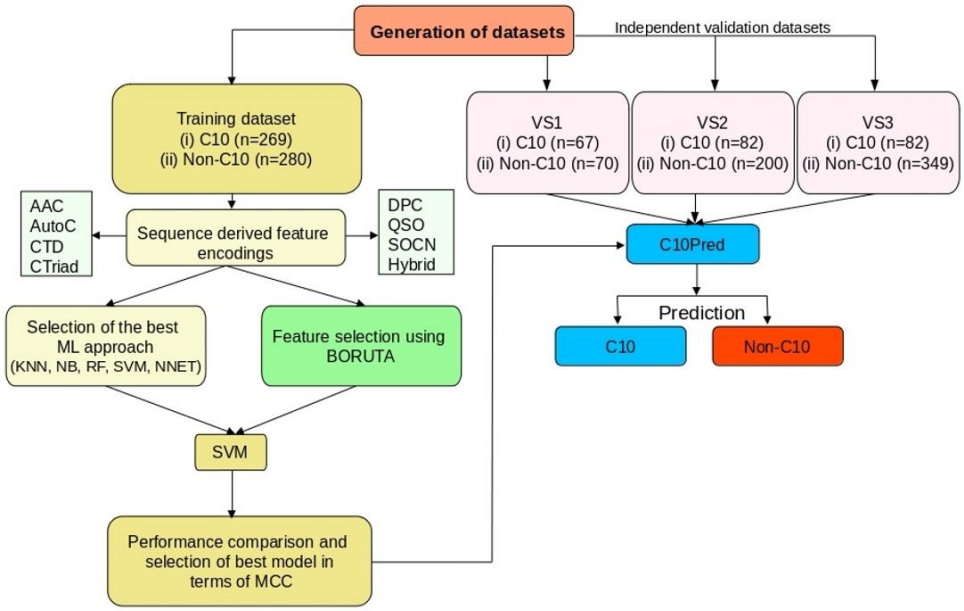

Figure 2. Schematic overview of the C10Pred tool demonstrating the four main stages of the predictor development. The first stage comprises the generation of the datasets, and the second stage consists of feature extraction from the primary amino acid sequences. In the third stage, we constructed five ML-based classifiers, namely, KNN, NB, RF, SVM, and NNET, using different feature sets and selected the best classifier. In parallel, we also performed feature selection using the Boruta algorithm. Finally, SVM was selected as the best ML classifier, and the performance of various optimal feature encodings was evaluated. KNN: K-nearest neighbors; NB: Naive Bayes; RF: random forest; SVM: support vector machines; NNET: neural network.

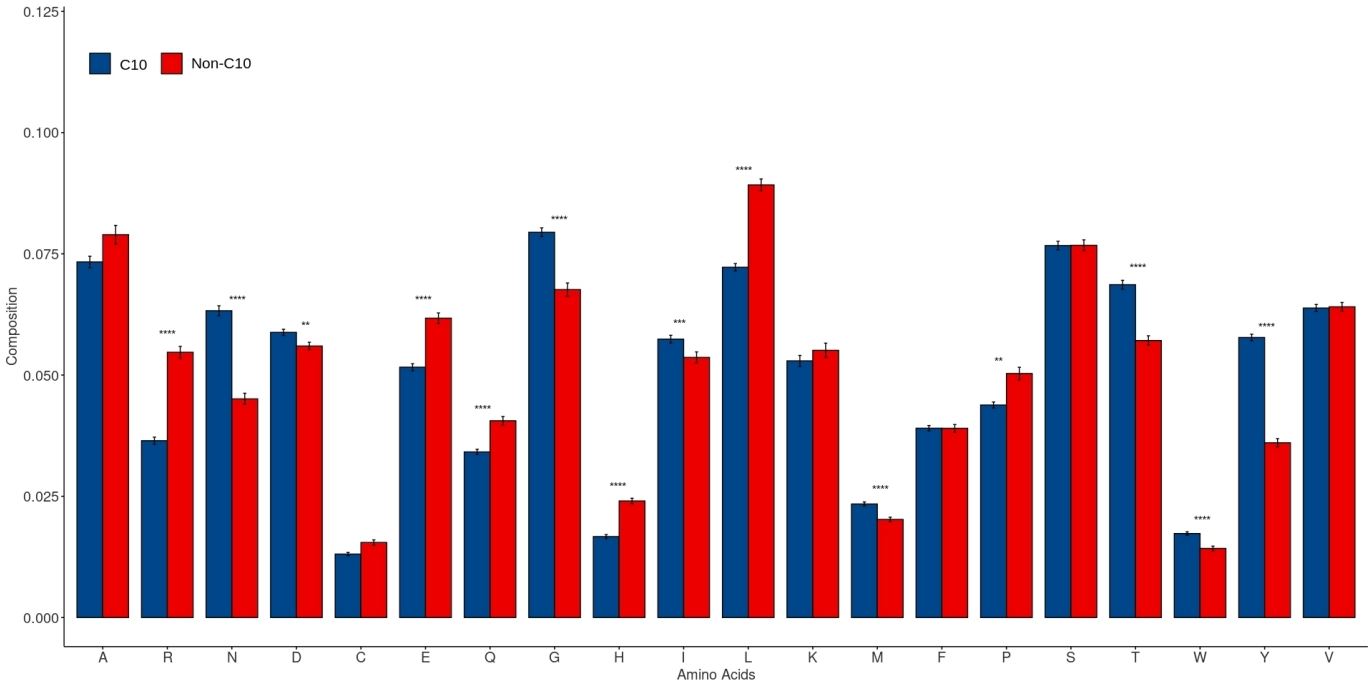

Figure 3. Average amino acid composition (AAC) difference between C10 and non-C10 family sequences. Asterisks above the bars indicate the *p*-value (** = $p < 0.01$; *** = $p < 0.001$; and **** = $p < 0.0001$).

2.4. Comparison of Various Machine Learning Classifiers

To assess the performance of various ML classifiers, we exploited five commonly used ML approaches (KNN, NB, RF, SVM, and NNET) on seven independent feature encodings and their hybrid. The performance of all these classifiers was assessed by using 10-fold cross-validation. Our comparative analysis suggests that the average performance of SVM was consistently better than four other classifiers in terms of accuracy and MCC on multiple feature encodings (Figures 4 and S2). Although the average performance of NNET was equally better, SVM demonstrated a slight edge by performing better on 4/8 descriptors (e.g., AAC, AutoC, CTD, and CTriad). In contrast, DPC, SOCN, and hybrid feature encodings performed better when NNET was used to train the model. Similar performance was observed for both these models when QSO was used as an input feature. These data indicated that SVM was the best performing classifier, and thus it was selected for further analysis.

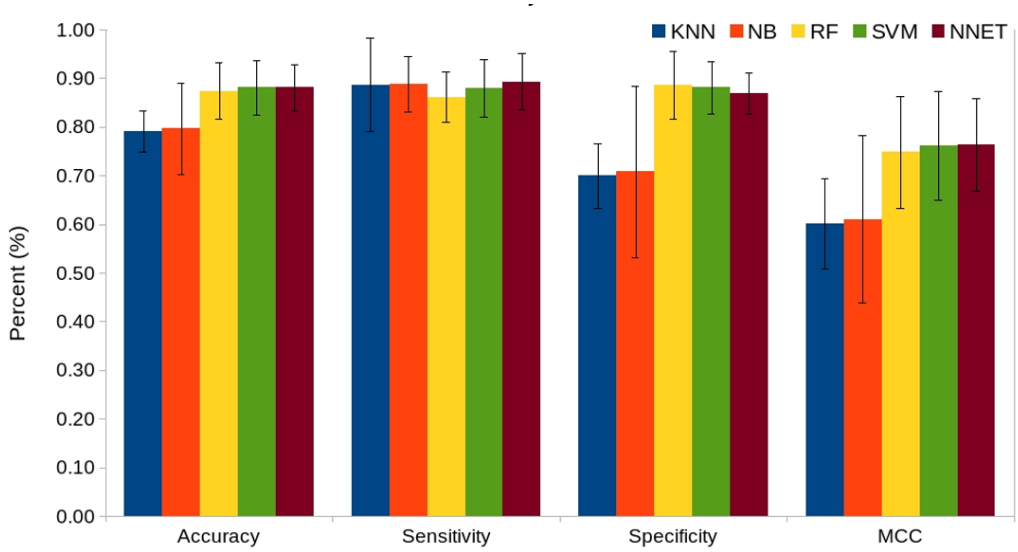

Figure 4. Average performance comparison of five ML-based classifiers (KNN, NB, RF, SVM, and NNET) using eight different feature encodings. The performance of each individual feature encoding for all classifiers is shown in Figure S2.

2.5. Performance Evaluation of Various Feature Encodings

We used SVM to probe the potential of each feature encoding in correctly differentiating C10 and non-C10 peptidases using 10-fold cross-validation. The performance achieved by each descriptor is shown in Table 2. Our data on the 10-fold cross-validation test shows that DPC followed by hybrid features achieved the best performance with accuracy scores of 93.4% and 92.9%, respectively. In addition to the high accuracy, these two features also exhibited encouraging MCC values that ranged between 0.85–0.86. In contrast, the accuracy scores for other descriptors (AAC, AutoC, CTD, CTriad, and QSO) were reasonable (84–90%), although with limited MCC scores. Models based on SOCN were the worst performing, with an accuracy of 75% and an MCC of 0.508.

Table 2. Performance of SVM on various feature encodings in 10-fold cross-validation. Accuracy scores ≥ 90% are in bold.

Features	Dimension Size	Accuracy	Sensitivity	Specificity	MCC
AAC	20	0.880	0.874	0.886	0.759
AutoC	720	0.883	0.877	0.889	0.767
CTD	147	**0.900**	0.892	0.907	0.800
CTriad	343	0.878	0.874	0.882	0.756
DPC	400	**0.934**	0.944	0.925	0.869
QSO	100	0.889	0.885	0.893	0.778
SOCN	60	0.754	0.755	0.754	0.508
Hybrid	1790	**0.929**	0.944	0.914	0.858

2.6. Optimal Feature Selection for Each Encoding

Recognizing that almost all the features except AAC and SOCN have large dimension sizes (≥100), some of the encodings might be superfluous or may not be equally significant. Therefore, this necessitates the application of feature selection protocol to eliminate redundant and insignificant encodings. We applied the Boruta algorithm to explore if it was able to slash the feature dimensions and affect the overall performance. Table 3 compares the performance achieved by various feature encodings using optimal features when classifying C10 and non-C10 peptidases. From this table, we also observed that when predicting C10 and non-C10 peptidases, the number of features was significantly reduced for hybrid features (92.23%), AutoC (85.83%), CTriad and DPC (~80%), QSO (55%), and CTD (39.45%). A limited number of features (3.33%) were removed for SOCN, while no dimension reduction was observed for AAC.

Table 3. The best performance achieved by various feature encodings using optimal features on 10-fold cross-validation. Values in bold indicate improvement in performance by at least 2% after feature selection.

Features	Dimension Size	Accuracy	Sensitivity	Specificity	MCC
AutoC	102	**0.900**	0.881	0.918	0.800
CTD	89	0.907	0.903	0.911	0.814
CTriad	67	0.883	0.874	0.893	0.767
DPC	79	0.925	0.926	0.925	0.851
QSO	45	**0.913**	0.907	0.918	0.825
SOCN	58	0.756	0.755	0.757	0.512
Hybrid	139	**0.956**	0.944	0.968	0.913

Following the reduction of feature dimensions, we explored the performance of each feature encoding using the optimal features and compared it with the respective controls (all features). Figure 5 shows a marginal improvement in the performance of most feature encodings, especially in the AutoC, QSO, and the hybrid, by 1.64%, 2.37%, and 2.73%, respectively, as compared to their controls. The improvement shown by CTD, CTriad, and SOCN is only marginal (<1%). Interestingly, there was a slight decrease of 0.01% in the performance of DPC when optimal feature sets were used.

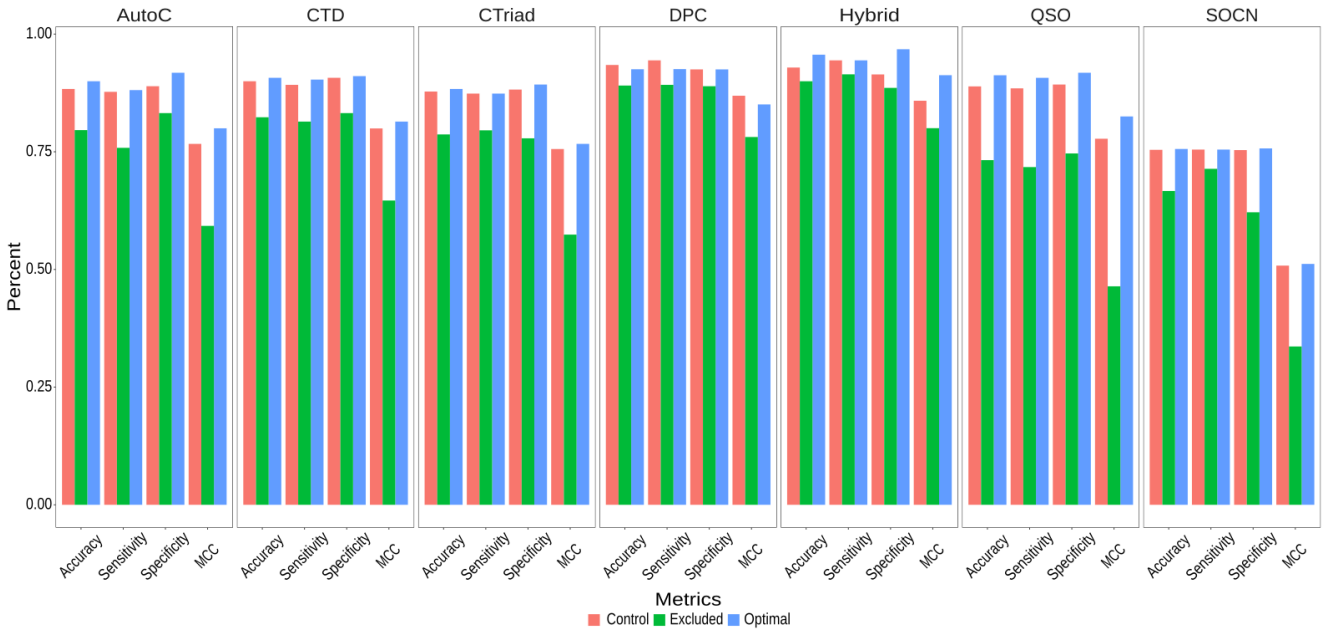

Figure 5. Performance comparison of SVM-based models using all features (control), excluded features and optimal features.

Next, to examine whether the optimal features are any better than the features excluded for each feature encodings, we developed prediction models based on excluded features and compared their performance with the control (using all features) and the optimal features. We observe that the models based on optimal features performed consistently better than those based on the excluded features (Figure 5). Notably, the average accuracy achieved by the optimal feature-based models is about 10% higher than the models based on excluded features and 1% higher compared to the controls when predicting C10 peptidases. Similarly, models based on optimal features exhibit better MCC scores than the control and excluded features. For example, using control feature encodings for AutoC, QSO, and hybrid, the classifier exhibited the MCC scores of 0.767, 0.778, and 0.858, respectively. However, using optimal features for these encodings, a significant increase in their MCC scores was observed (e.g., 0.80, 0.825, and 0.913). In contrast, classifiers based on excluded features performed worst, and the MCC scores for these three encodings are 0.593, 0.464, and 0.80, respectively. These data suggest that the Boruta algorithm identified important features contributing to improved performance and overall dimension reduction.

2.7. Performance Comparison on Independent Datasets

It is well known that the testing of an ML algorithm on the training data does not provide the best clue regarding its performance on the unseen data because of the deceivingly overall high accuracies [42]. Therefore, to verify whether the consistence performance is shown by various feature encodings, we assessed each of these optimal feature-based encodings on the independent validation set VS1. The results on VS1 indicate that hybrid features show the best performance, which is similar and consistent with the performance obtained in the 10-fold cross-validation test (Tables 3 and 4). This hybrid model using optimal features achieves an accuracy, sensitivity, specificity, and MCC of 0.927, 0.896, 0.957, and 0.855, respectively. Specifically, the accuracy and MCC achieved using hybrid encodings are approximately 3–19% and 7–38% higher, respectively, than the other feature encodings. Furthermore, to graphically visualize the performance of various encodings, an ROC curve was generated by computing and plotting the true positive rate (TPR) versus the false positive rate (FPR) (Figure 6). In such plots, a higher AUC score indicates a better classifier performance. From this figure, we again observe that the hybrid classifier using optimal features showed the best AUC of 0.98. These data demonstrate that the hybrid

model using optimal features has the potential to accomplish promising performance. Therefore, this classifier was selected as a final model. Although DPC based classifier also exhibited a similar AUC value (Figure 6), it showed poor performance when other evaluation metrics such as accuracy and MCC were considered.

Table 4. Performance comparison of various optimal feature encodings on VS1 dataset. Accuracy scores $\geq$ 90% are in bold.

Features	Accuracy	Sensitivity	Specificity	MCC
AutoC	0.839	0.731	0.943	0.692
CTD	0.839	0.791	0.886	0.681
CTriad	0.861	0.776	0.943	0.731
DPC	0.891	0.836	0.943	0.785
QSO	0.869	0.836	0.900	0.738
SOCN	0.737	0.701	0.771	0.474
Hybrid	**0.927**	0.896	0.957	0.855

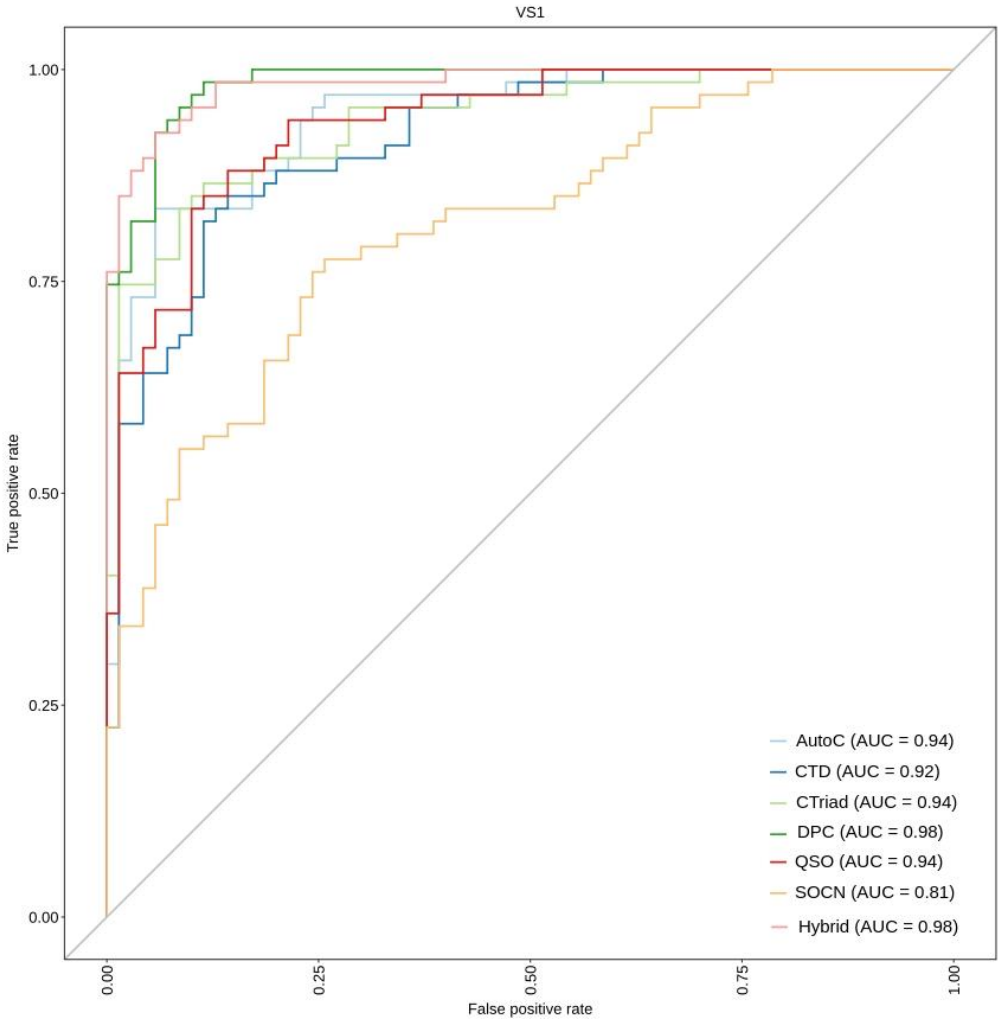

Figure 6. Comparison of binormal receiver operating characteristics (ROC) curves for various prediction models on the independent dataset VS1. Higher scores indicate better performance of that specific model.

To further assess the performance of C10Pred we used an additional validation set, VS2. Figure 7 shows the confusion matrix for predicting the binary classification of C10 peptidases. Specifically, the figure shows that 4/82 positive sequences were incorrectly

predicted, whereas only 3/200 negative sequences were classified as positive sequences. The three negative sequences, incorrectly predicted as positive sequences, include papain domain containing C1 family cysteine protease, caspase P20 domain-containing protein, and a C25 family cysteine protease, respectively. Interestingly, similar to the amino acid composition of the positive dataset (Figure 3), all these three sequences also exhibited a higher percentage of TYR residue than the average value for the negative sequences. Similarly, the $\frac{3}{4}$ positive sequences, wrongly identified as negative sequences, exhibited a TYR profile similar to the negative sequences. The amino acid composition of some other residues (e.g., TRP, LEU, and GLY) in these four positive sequences also deviated and was analogous to their counterparts in the negative dataset. The accuracy, MCC, and AUC achieved by our proposed method on the VS2 dataset are 0.975, 0.94, and 0.968, respectively. As mentioned in the methods section, the negative dataset comprises sequences representing all other cysteine peptidases except C10 family proteases. Therefore, to assess how the model behaves if a more diverse set of sequences is used as a negative dataset rather than just other families of cysteine proteases, we compiled another non-redundant dataset of 349 negative sequences from the UNIPROT [43] database. This new set of negative sequences was merged with the positive sequences of the VS2 dataset to form additional validation set VS3 (Table 1). Compared to VS1 and VS2, VS3 consists of all bacterial sequences except the C10 family or streptopain proteins. It should be noted that both VS2 and VS3 validation sets comprise sequences that show <50% sequence similarity with the positive data. On assessing the performance of C10Pred using the VS3 dataset, a slight increase in accuracy (0.979%), with a small decrease in MCC (0.933), was observed (Figure S3). Altogether, the results demonstrate the remarkable performance achieved by C10Pred, which could be further enhanced by exploiting large-scale training data when it becomes available in the future.

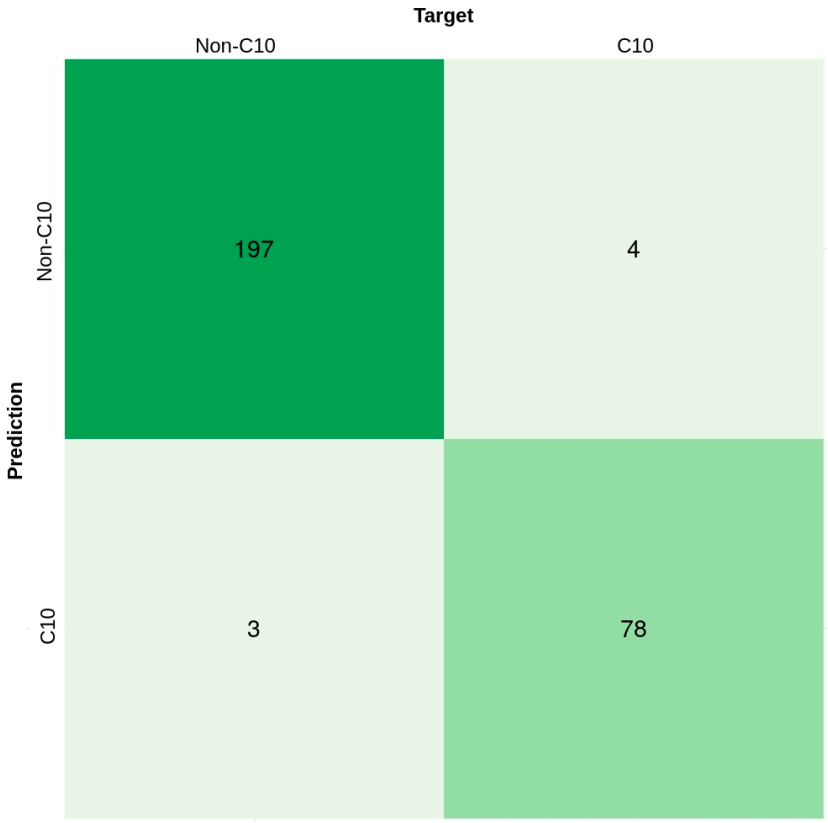

Figure 7. Confusion matrix of predicted results on additional independent dataset VS2. The matrix represents the output distribution for each of the two classes (C10 or non-C10).

2.8. Software Availability

To make our method publicly available so that potential users may benefit from it, both the standalone as well as the webserver version for C10Pred are freely accessible at the following link: https://procarb.org/c10pred/ (accessed 16 August 2022). The input to both the versions is the fasta formatted sequences, and the prediction results are available as a downloadable comma-separated file (CSV). All instructions and datasets used in this work are available on the C10Pred homepage.

3. Discussion

S. pyogenes expresses a highly conserved virulence factor streptopain (a C10 family cysteine peptidase) known to degrade an array of GAS and host proteins [44]. Although much of the work on this C10 cysteine protease has been reported from *Streptococcal* strains, many recent studies have identified this protease or its homologs in other bacterial species [34,35,45,46]. The biological activities and molecular functions of proteins can be predicted from their amino acid sequences [47]. Therefore, in the present study, we exploited the available C10 family sequences to develop the ML-based predictor, C10Pred, to classify the C10 family proteases.

We utilized seven feature encodings (AAC, AutoC, CTD, CTriad, DPC, QSO, and SOCN) and also combined all these features (hybrid) to predict C10 proteases using SVM. These features have been used extensively in various sequence-based protein classification problems [48–52]. The performance results from each encoding performed well, especially with DPC and CTD-based classifiers. However, the predictor's performance using SOCN descriptors was moderate. Interestingly, when we applied the feature selection protocol, we observed that optimal hybrid encodings outperformed the other features. Therefore, we considered it to be the most efficient feature for the prediction of C10 enzymes. This dimension reduced the optimal hybrid feature set of 139 descriptors showing a >2% increase in the overall accuracy of the classifier and about a 5% increase in its MCC score. The corresponding sensitivity and specificity values for these optimal 139 features are 0.944 and 0.968, respectively. Selection of optimal features is one of the essential steps in developing ML-based models because the original set may contain redundant and non-informative features [49]. These copious non-informative and redundant descriptors, especially in the case of high dimensional features, affect the prediction accuracy. Therefore, selecting optimal features is regarded as one of the most influential steps in ML-based prediction [53–56]. Recognizing the potential of feature selection, we applied the Boruta feature selection method, which has been widely applied effectively in several biological applications [57–59], and consequently identified optimal features. Among these, the major contribution was from AutoC (~27%), followed by CTD, DPC, QSO, CTriad, AAC, and SOCN. Although AutoC descriptors were the major contributors, the top 10 most important features were dominated by DPC (5/10), QSO (4/10), and a single Y residue of AAC. These top-scoring dipeptides included GW, GC, WG, GY, and YN. It should be noted that these dipeptides comprise residues that are more abundant in C10 peptidases than non-C10 peptidases (Figure 3).

Based on the performance obtained on the hybrid model using an optimal feature set, the SVM-based predictor C10Pred was constructed. Moreover, the dataset generated in this work has a stringent sequence identity of $\leq$40%, which is essential to avoid overestimating the predictive performance of a predictor [49]. Importantly, this is the first ML-based method for predicting C10 family peptidases using sequence-derived information, and is freely available as a web server. Since there is no other method available for the prediction of C10 peptidases or their homologs, a direct comparison is impossible. Although C10Pred exhibited acceptable predictive performance, there is scope for further improvements. For example, constructing a model on a larger dataset when it becomes available, testing other feature encodings, and exploiting different ML algorithms such as stochastic gradient boosting [60].

4. Materials and Methods

4.1. Data Acquisition and Data Organization

All protein sequences representing "the Peptidase_C10" family (PFAM ID: PF01640) within the PFAM [61] database were retrieved. All non-standard amino acids containing sequences were removed, and sequences shorter than 100 amino acids were also excluded. The remaining sequences were subjected to a redundancy removal by applying CD-HIT v4.8.1 [62] with the 40% sequence identity cut-off.

The negative dataset was generated as follows: (i) retrieved all PFAM sequences belonging to various cysteine peptidase clans/families except the family C10 which was used as a positive dataset. (ii) Non-standard amino acids containing sequences were eliminated. (iii) Sequences having a length between 100 and 2300 amino acids were retained only, and (iv) We further filtered the negative dataset at 40% sequence identity cut-off using CD-HIT. These steps generated a large number of over 47,000 sequences in the negative dataset. To generate a balanced dataset, we randomly selected negative samples that were similar in number to the positive dataset. To ensure a limited similarity between the positive and negative datasets, we removed all the negative samples that showed $\geq$25% sequence identity with the positive dataset.

Both these datasets mentioned above were combined and divided into a training and an independent validation set (VS1) by using the createDataPartition function of the CARET (short for Classification And REgression Training) package [63] available in R (https://www.r-project.org/: accessed 16 August 2022).

Furthermore, to assess the robustness of our method, we constructed an additional independent validation set (VS2) by retrieving all the streptopain sequences available in the NCBI protein database. After filtering non-standard amino acid-containing sequences, the sequences were further processed for redundancy removal at a 50% sequence identity cut-off. Subsequently, we removed all the sequences that shared $\geq$50% sequence similarity with the positive dataset. Again, a negative dataset of 200 sequences was randomly constructed and combined with these sequences. The overall summary of the datasets is provided in Table 1.

4.2. Feature Encoding

To develop an ML model, sequences with varying lengths were converted to fixed-length feature vectors using feature encoding algorithms. In this work, we used an R package 'protr' [64] to generate seven different features that have been extensively used in previous works. These features represent major compositional and physicochemical characteristics of a sequence and are described below:

4.2.1. Amino Acid Composition (AAC)

The AAC of a protein sequence represents the fraction of each of the 20 standard amino acid residues. AAC has a fixed length of 20 features, and it can be mathematically represented as

$$AAC(i) = \frac{AA_i}{K} \quad (1)$$

where AA_i represents the number of amino acids of type i, and K denotes protein sequence length.

4.2.2. Autocorrelation (AutoC)

Autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence. AutoC descriptors are grouped into three types: (i) Moran, (ii) Moreau-Broto, and (iii) Geary, and can be denoted by Equations (2)–(4), respectively.

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} d = 1, 2, \ldots, nlag \quad (2)$$

where d is the lag of autocorrelation; P_i and P_{i+d} are the amino acid properties at position i and i+d; nlag represents the maximum value of the lag.

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}\left(P_i - \overline{P}\right)\left(P_{i+d} - \overline{P}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} d = 1, 2, \ldots, 30 \quad (3)$$

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}\left(P_i - P_{i+d}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} d = 1, 2, \ldots, 30 \quad (4)$$

where d is the autocorrelation lag, P_i and P_{i+d} are the amino acid properties at positions i and i+d, and $\overline{P}$ is the average value of property P denoted as: $\overline{P} = \sum_{i=1}^{N} P_i/N$.

4.2.3. Composition (C), Transition (T), and Distribution (D) (CTD)

The CTD descriptors were described more than two decades ago to predict protein folding classes and represent the distribution of amino acid patterns for specific structural and physicochemical properties of protein sequences [65,66]. In CTD, the 20 standard amino acids are divided into three groups based on seven different types of physicochemical properties such as hydrophobicity, normalized van der Waals volume, polarizability, polarity, etc. (Table S2). In CTD, C is the fraction of polar, neutral, and hydrophobic residues of a given protein sequence:

$$C(a) = \frac{Z_a}{K}, a \in \{neutral, polar, hydrophobic\} \quad (5)$$

Za is the number of amino acids of type a in the given sequence.

T computes the percentage frequency of a specific property of an amino acid progressed by another property:

$$T(ab) = \frac{Z_{ab} + Z_{ba}}{K-1}, a, b \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\} \quad (6)$$

where Z_{ab} and Z_{ba} represent the number of dipeptides encoded as ab and ba in the sequence.

Finally, D comprises five values for each of the three groups and determines the percentage of a target sequence length within which 25, 50, 75, and 100% of the amino acids of a specific property are located. In summary, CTD generates a feature vector of 147 dimensions.

4.2.4. Conjoint Triad (CTriad)

The CTriad encodings were first used to predict protein-protein interactions [67]. In CTriad, a protein sequence is depicted as a vector space containing features of amino acids. Consequently, the vector space is trimmed by clustering the 20 naturally occurring amino acids based on their dipoles and side chains volumes, resulting in a 343-dimensional feature vector for any given protein sequence.

4.2.5. Dipeptide Composition (DPC)

DPC is a fixed length of 400 (20 × 20) features and is defined as the frequency of two amino acid types in a given protein sequence:

$$DPC(ab) = \frac{z_{ab}}{K-1} \quad (7)$$

4.2.6. Quasi-Sequence Order (QSO)

QSO descriptors are derived by measuring the physicochemical distance between the amino acids of a given protein sequence and result in a fixed length of a 100-dimensional feature vector [68,69]. The first 20 quasi-sequence-order descriptors are defined as:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{maxlag} \tau d} \quad \mathrm{r} = 1, 2, \ldots, 20 \tag{8}$$

where f_r is the normalized occurrence for amino acid type, I and w is a weighting factor (w = 0.1). The other 30 quasi-sequence-order are defined as:

$$X_d = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{maxlag} \tau d} \quad \mathrm{d} = 21, 22, \ldots, 30 \tag{9}$$

4.2.7. Sequence Order Coupling Number (SOCN)

The d-th rank sequence-order-coupling number is defined as:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 d = 1, 2, \ldots, maxlag \tag{10}$$

where $d_{i,i+d}$ is the maximum lag, and the protein length must not be less than max lag.

4.3. Machine Learning Models

To get a quick approximation of the best ML classifier, we assessed five commonly used ML algorithms, namely, K-nearest neighbors (KNN), naive Bayes (NB), random forest (RF), support vector machines (SVM), and neural network (NNET), by using CARET [63]. Using the 10-fold cross-validation (CV) approach, we assessed the performance of a given set of eight feature encodings (AAC, AutoC, CTD, CTriad, DPC, QSO, SOCN, and Hybrid) using default parameters for each corresponding ML algorithm.

4.4. Feature Selection

To improve the feature representation capability and determine the subset of ideal features that can correctly classify C10 peptidase (streptopain) and non-C10 peptidase (non-streptopain) sequences, we used the R implementation of the Boruta package (v7.0.0) [70]. Boruta is a feature selection algorithm and feature ranking based on the RF algorithm. Boruta analyzes the feature importance values calculated for the real predictor variables against the shadow variables (i.e., variables created by the permutation of these variables across observations). For each run, an RF is trained using a double length set of predictor variables comprising of an equivalent number of actual and shadow variables. For each of the real predictor variables, a statistical test is performed comparing its significance in relation to the utmost importance value accomplished by a shadow variable. Each variable can be classified as important or unimportant based on the importance values. Finally, all unimportant and shadow variables are eliminated. The process is repeated until all variables have been classified as important or unimportant, or a specific number of runs (maxRuns) have been achieved [71]. The default value of the maxRuns parameter is 100, and we observed that it was too small for the algorithm to classify variables as important or unimportant. Therefore, we set the max runs parameter to 1000. Any remaining tentative features were checked by the TentativeRoughFix function, which compares the median Z-score of a tentative feature and the median of maximum Z-scores among shadow features (MZSF) across the previous RF runs and eventually makes a decision. Overall, Boruta performs a top-down approach for relevant features by comparing the set of original attributes [72] and has been used in many feature selection tasks [71,73,74].

4.5. Performance Evaluation Metrics

To estimate the performance of our ML models, we used four widely used metrics that estimate the quality of binary classification. These include sensitivity, specificity, accuracy, and Matthews' correlation coefficient (MCC) and are expressed as:

$$Sensitivity = \frac{TP}{TP + TN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where *TP*, *TN*, *FP*, and *FN* represent the true positive, true negative, false positive, and false negative, respectively. In all cases, the higher the value, the better the prediction performance.

5. Conclusions

Cysteine peptidases that belong to the C10 family are represented by streptopain or streptopain-like proteases. These enzymes are critical virulence factors that cause tissue damage and severe lethal effect in GAS-infected mice, involved in toxic shock syndrome and apoptosis. Initially identified in all GAS, this protease has been identified in several other bacterial species. Therefore, an attempt was made to construct a novel ML model (C10Pred) using SVM and optimal features from the primary amino acid sequences. The predictive performance of C10Pred on 10-fold cross-validation and three independent datasets (VS1, VS2, and VS3) exhibited encouraging performance. Our predictor is a handy tool to classify novel C10 family or streptopain-like proteins, and offers essential information for researchers interested in C10 family proteases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23179518/s1.

Author Contributions: Conceptualization, A.M. and C.-B.K.; software, A.M.; validation, T.A.D. and C.-B.K.; formal analysis, A.M. and N.M; resources, T.A.D. and C.-B.K.; writing—original draft preparation, A.M., N.M. and T.A.D.; writing—review and editing, A.M. and C.-B.K.; supervision, A.M. and C.-B.K.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (2021R1I1A1A01056363).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available within this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cannon, J.W.; Zhung, J.; Bennett, J.; Moreland, N.J.; Baker, M.G.; Geelhoed, E.; Fraser, J.; Carapetis, J.R.; Jack, S. The economic and health burdens of diseases caused by group A Streptococcus in New Zealand. *Int. J. Infect. Dis.* **2021**, *103*, 176–181. [CrossRef] [PubMed]
2. Nelson, G.E.; Pondo, T.; Toews, K.-A.; Farley, M.M.; Lindegren, M.L.; Lynfield, R.; Aragon, D.; Zansky, S.M.; Watt, J.P.; Cieslak, P.R.; et al. Epidemiology of Invasive Group A Streptococcal Infections in the United States, 2005–2012. *Clin. Infect. Dis.* **2016**, *63*, 478–486. [CrossRef] [PubMed]
3. Carapetis, J.R.; Steer, A.C.; Mulholland, E.K.; Weber, M. The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* **2005**, *5*, 685–694. [CrossRef]

4. Bowen, A.C.; Tong, S.Y.C.; Chatfield, M.D.; Carapetis, J.R. The microbiology of impetigo in Indigenous children: Associations between Streptococcus pyogenes, Staphylococcus aureus, scabies, and nasal carriage. *BMC Infect. Dis.* **2014**, *14*, 727. [CrossRef]
5. Bowen, A.C.; Mahé, A.; Hay, R.J.; Andrews, R.M.; Steer, A.C.; Tong, S.Y.C.; Carapetis, J.R. The Global Epidemiology of Impetigo: A Systematic Review of the Population Prevalence of Impetigo and Pyoderma. *PLoS ONE* **2015**, *10*, e0136789. [CrossRef]
6. Cunningham, M.W. Pathogenesis of Group A Streptococcal Infections. *Clin. Microbiol. Rev.* **2000**, *13*, 470–511. [CrossRef]
7. Castro, S.A.; Dorfmueller, H.C. A brief review on Group A Streptococcus pathogenesis and vaccine development. *R. Soc. Open Sci.* **2021**, *8*, 201991. [CrossRef]
8. Carapetis, J.R.; Beaton, A.; Cunningham, M.W.; Guilherme, L.; Karthikeyan, G.; Mayosi, B.M.; Sable, C.; Steer, A.; Wilson, N.; Wyber, R.; et al. Acute rheumatic fever and rheumatic heart disease. *Nat. Rev. Dis. Prim.* **2016**, *2*, 15084. [CrossRef]
9. Hoy, W.E.; White, A.V.; Dowling, A.; Sharma, S.K.; Bloomfield, H.; Tipiloura, B.T.; Swanson, C.E.; Mathews, J.D.; McCredie, D.A. Post-streptococcal glomerulonephritis is a strong risk factor for chronic kidney disease in later life. *Kidney Int.* **2012**, *81*, 1026–1032. [CrossRef]
10. Marshall, C.S.; Cheng, A.C.; Markey, P.G.; Towers, R.J.; Richardson, L.J.; Fagan, P.K.; Scott, L.; Krause, V.L.; Currie, B.J. Acute Post-Streptococcal Glomerulonephritis in the Northern Territory of Australia: A Review of 16 Years Data and Comparison with the Literature. *Am. J. Trop. Med. Hyg.* **2011**, *85*, 703–710. [CrossRef]
11. Oliver, J.; Pierse, N.; Williamson, D.A.; Baker, M.G. Estimating the likely true changes in rheumatic fever incidence using two data sources. *Epidemiol. Infect.* **2017**, *146*, 265–275. [CrossRef] [PubMed]
12. Vogel, A.M.; Lennon, D.R.; van der Werf, B.; Diack, M.; Neutze, J.M.; Horsfall, M.; Emery, D.; Wong, W. Post-streptococcal glomerulonephritis: Some reduction in a disease of disparities. *J. Paediatr. Child Health* **2018**, *55*, 652–658. [CrossRef] [PubMed]
13. Stockmann, C.; Ampofo, K.; Hersh, A.L.; Blaschke, A.J.; Kendall, B.A.; Korgenski, K.; Daly, J.; Hill, H.R.; Byington, C.L.; Pavia, A.T. Evolving Epidemiologic Characteristics of Invasive Group A Streptococcal Disease in Utah, 2002–2010. *Clin. Infect. Dis.* **2012**, *55*, 479–487. [CrossRef] [PubMed]
14. Lamagni, T.L.; Efstratiou, A.; Vuopio-Varkila, J.; Jasir, A.; Schalén, C.; Euro, S. The epidemiology of severe Streptococcus pyogenes associated disease in Europe. *Eurosurveillance* **2005**, *10*, 9–10. [CrossRef]
15. Watkins, D.A.; Johnson, C.O.; Colquhoun, S.M.; Karthikeyan, G.; Beaton, A.; Bukhman, G.; Forouzanfar, M.H.; Longenecker, C.T.; Mayosi, B.M.; Mensah, G.A.; et al. Global, Regional, and National Burden of Rheumatic Heart Disease, 1990–2015. *N. Engl. J. Med.* **2017**, *377*, 713–722. [CrossRef]
16. Gubba, S.; Low, D.E.; Musser, J.M. Expression and Characterization of Group A *Streptococcus* Extracellular Cysteine Protease Recombinant Mutant Proteins and Documentation of Seroconversion during Human Invasive Disease Episodes. *Infect. Immun.* **1998**, *66*, 765–770. [CrossRef]
17. Gerlach, D.; Knöll, H.; Köhler, W.; Ozegowski, J.-H.; Hríbalova, V. Isolation and characterization of erythrogenic toxins V. Communication: Identity of erythrogenic toxin type B and Streptococcal proteinase precursor. *Zentralbl. Bakteriol. Mikrobiol. Hyg. A Med. Mikrobiol. Infekt. Parasitol.* **1983**, *255*, 221–233. [CrossRef]
18. Hauser, A.R.; Schlievert, P.M. Nucleotide sequence of the streptococcal pyrogenic exotoxin type B gene and relationship between the toxin and the streptococcal proteinase precursor. *J. Bacteriol.* **1990**, *172*, 4536–4542. [CrossRef]
19. Rawlings, N.D.; Barrett, A.J.; Thomas, P.D.; Huang, X.; Bateman, A.; Finn, R.D. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **2018**, *46*, D624–D632. [CrossRef]
20. Rawlings, N.D.; Barrett, A.J. Introduction: The Clans and Families of Cysteine Peptidases. In *Handbook of Proteolytic Enzymes*; Academic Press: Cambridge, MA, USA, 2013; pp. 1743–1773. [CrossRef]
21. Chen, C.-Y.; Luo, S.-C.; Kuo, C.-F.; Lin, Y.-S.; Wu, J.-J.; Lin, M.T.; Liu, C.-C.; Jeng, W.-Y.; Chuang, W.-J. Maturation Processing and Characterization of Streptopain. *J. Biol. Chem.* **2003**, *278*, 17336–17343. [CrossRef]
22. Kagawa, T.F.; Cooney, J.C.; Baker, H.M.; McSweeney, S.; Liu, M.; Gubba, S.; Musser, J.M.; Baker, E.N. Crystal structure of the zymogen form of the group A *Streptococcus* virulence factor SpeB: An integrin-binding cysteine protease. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2235–2240. [CrossRef] [PubMed]
23. Musser, J.M.; Hauser, A.R.; Kim, M.H.; Schlievert, P.M.; Nelson, K.; Selander, R.K. Streptococcus pyogenes causing toxic-shock-like syndrome and other invasive diseases: Clonal diversity and pyrogenic exotoxin expression. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2668–2672. [CrossRef] [PubMed]
24. Liu, T.Y.; Elliott, S.D. Streptococcal proteinase: The zymogen to enzyme transfromation. *J. Biol. Chem.* **1965**, *240*, 1138–1142. [CrossRef]
25. Liu, T.Y.; Elliott, S.D. Activation of Streptococcal Proteinase and its Zymogen by Bacterial Cell Walls. *Nature* **1965**, *206*, 33–34. [CrossRef]
26. Wang, C.-C.; Houng, H.-C.; Chen, C.-L.; Wang, P.-J.; Kuo, C.-F.; Lin, Y.-S.; Wu, J.-J.; Lin, M.T.; Liu, C.-C.; Huang, W.; et al. Solution structure and backbone dynamics of streptopain: Insight into diverse substrate specificity. *J. Biol. Chem.* **2009**, *284*, 10957–10967. [CrossRef]
27. Walker, M.; Hollands, A.; Sanderson-Smith, M.; Cole, J.N.; Kirk, J.K.; Henningham, A.; McArthur, J.D.; Dinkla, K.; Aziz, R.; Kansal, R.G.; et al. DNase Sda1 provides selection pressure for a switch to invasive group A streptococcal infection. *Nat. Med.* **2007**, *13*, 981–985. [CrossRef]

28. Kapur, V.; Topouzis, S.; Majesky, M.W.; Li, L.-L.; Hamrick, M.R.; Hamill, R.J.; Patti, J.M.; Musser, J.M. A conserved Streptococcus pyogenes extracellular cysteine protease cleaves human fibronectin and degrades vitronectin. *Microb. Pathog.* **1993**, *15*, 327–346. [CrossRef]
29. Wu, G.; Mahajan, N.; Dhawan, V. Acknowledged Signatures of Matrix Metalloproteinases in Takayasu's Arteritis. *BioMed Res. Int.* **2014**, *2014*, 827105. [CrossRef]
30. Tamura, F.; Nakagawa, R.; Akuta, T.; Okamoto, S.; Hamada, S.; Maeda, H.; Kawabata, S.; Akaike, T. Proapoptotic Effect of Proteolytic Activation of Matrix Metalloproteinases by *Streptococcus pyogenes* Thiol Proteinase (*Streptococcus* Pyrogenic Exotoxin B). *Infect. Immun.* **2004**, *72*, 4836–4847. [CrossRef]
31. Stockbauer, K.E.; Magoun, L.; Liu, M.; Burns, E.H.; Gubba, S.; Renish, S.; Pan, X.; Bodary, S.C.; Baker, E.; Coburn, J.; et al. A natural variant of the cysteine protease virulence factor of group A Streptococcus with an arginine-glycine-aspartic acid (RGD) motif preferentially binds human integrins alphavbeta3 and alphaIIbbeta3. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 242–247. [CrossRef]
32. Byrne, D.P.; Wawrzonek, K.; Jaworska, A.; Birss, A.J.; Potempa, J.; Smalley, J.W. Role of the cysteine protease interpain A of *Prevotella intermedia* in breakdown and release of haem from haemoglobin. *Biochem. J.* **2009**, *425*, 257–264. [CrossRef] [PubMed]
33. Nelson, D.; Potempa, J.; Kordula, T.; Travis, J. Purification and characterization of a novel cysteine proteinase (periodontain) from Porphyromonas gingivalis. Evidence for a role in the inactivation of human alpha1-proteinase inhibitor. *J. Biol. Chem.* **1999**, *274*, 12245–12251. [CrossRef] [PubMed]
34. Gómez, E.; Alvarez, B.; Duchaud, E.; Guijarro, J.A. Development of a Markerless Deletion System for the Fish-Pathogenic Bacterium Flavobacterium psychrophilum. *PLoS ONE* **2015**, *10*, e0117969. [CrossRef] [PubMed]
35. Pérez-Pascual, D.; Lunazzi, A.; Magdelenat, G.; Rouy, Z.; Roulet, A.; Lopez-Roques, C.; Larocque, R.; Barbeyron, T.; Gobet, A.; Michel, G.; et al. The Complete Genome Sequence of the Fish Pathogen Tenacibaculum maritimum Provides Insights into Virulence Mechanisms. *Front. Microbiol.* **2017**, *8*, 1542. [CrossRef]
36. Tett, A.; Huang, K.D.; Asnicar, F.; Fehlner-Peach, H.; Pasolli, E.; Karcher, N.; Armanini, F.; Manghi, P.; Bonham, K.; Zolfo, M.; et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* **2019**, *26*, 666–679.e7. [CrossRef]
37. Ibrahim, M.; Subramanian, A.; Anishetty, S. Comparative pan genome analysis of oral Prevotella species implicated in periodontitis. *Funct. Integr. Genom.* **2017**, *17*, 513–536. [CrossRef]
38. Patra, A.K.; Yu, Z. Genomic Insights into the Distribution of Peptidases and Proteolytic Capacity among *Prevotella* and *Paraprevotella* Species. *Microbiol. Spectr.* **2022**, *10*. [CrossRef]
39. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
40. Potter, S.C.; Luciani, A.; Eddy, S.R.; Park, Y.; López, R.; Finn, R.D. HMMER web server: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W200–W204. [CrossRef]
41. Cantalapiedra, C.P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **2021**, *38*, 5825–5829. [CrossRef]
42. Štambuk, N.; Konjevoda, P. The Role of Independent Test Set in Modeling of Protein Folding Kinetics. *Adv. Exp. Med. Biol.* **2011**, *696*, 279–284. [CrossRef] [PubMed]
43. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef] [PubMed]
44. Carroll, R.K.; Musser, J.M. From transcription to activation: How group A streptococcus, the flesh-eating pathogen, regulates SpeB cysteine protease production. *Mol. Microbiol.* **2011**, *81*, 588–601. [CrossRef] [PubMed]
45. Bridel, S.; Bourgeon, F.; Marie, A.; Saulnier, D.; Pasek, S.; Nicolas, P.; Bernardet, J.-F.; Duchaud, E. Genetic diversity and population structure of Tenacibaculum maritimum, a serious bacterial pathogen of marine fish: From genome comparisons to high throughput MALDI-TOF typing. *Vet. Res.* **2020**, *51*, 60. [CrossRef]
46. Lithgow, K.V.; Buchholz, V.C.H.; Ku, E.; Konschuh, S.; D'Aubeterre, A.; Sycuro, L.K. Protease activities of vaginal Porphyromonas species disrupt coagulation and extracellular matrix in the cervicovaginal niche. *NPJ Biofilms Microbiomes* **2022**, *8*, 8. [CrossRef]
47. Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681–697. [CrossRef]
48. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J. Proteome Res.* **2018**, *17*, 2715–2726. [CrossRef]
49. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.* **2018**, *9*, 476. [CrossRef]
50. Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.-C. mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. *Int. J. Mol. Sci.* **2019**, *20*, 1964. [CrossRef]
51. Singh, O.; Hsu, W.-L.; Su, E.C.-Y. ILeukin10Pred: A Computational Approach for Predicting IL-10-Inducing Immunosuppressive Peptides Using Combinations of Amino Acid Global Features. *Biology* **2021**, *11*, 5. [CrossRef]
52. Malik, A.; Subramaniyam, S.; Kim, C.-B.; Manavalan, B. SortPred: The first machine learning based predictor to identify bacterial sortases and their classes using sequence-derived information. *Comput. Struct. Biotechnol. J.* **2021**, *20*, 165–174. [CrossRef] [PubMed]

53. Wang, M.; Zhao, X.-M.; Takemoto, K.; Xu, H.; Li, Y.; Akutsu, T.; Song, J. FunSAV: Predicting the Functional Effect of Single Amino Acid Variants Using a Two-Stage Random Forest Model. *PLoS ONE* **2012**, *7*, e43847. [CrossRef] [PubMed]
54. Song, J.; Wang, H.; Wang, J.; Leier, A.; Marquez-Lago, T.; Yang, B.; Zhang, Z.; Akutsu, T.; Webb, G.I.; Daly, R.J. PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Rep.* **2017**, *7*, 6862. [CrossRef] [PubMed]
55. Wei, L.; He, W.; Malik, A.; Su, R.; Cui, L.; Manavalan, B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Briefings Bioinform.* **2021**, *22*, bbaa275. [CrossRef] [PubMed]
56. Basith, S.; Lee, G.; Manavalan, B. STALLION: A stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Briefings Bioinform.* **2022**, *23*, bbab376. [CrossRef]
57. Li, Z.; Guo, W.; Ding, S.; Chen, L.; Feng, K.; Huang, T.; Cai, Y.-D. Identifying Key MicroRNA Signatures for Neurodegenerative Diseases With Machine Learning Methods. *Front. Genet.* **2022**, *13*, 880997. [CrossRef]
58. Uchida, Y.; Yoshida, S.; Arita, Y.; Shimoda, H.; Kimura, K.; Yamada, I.; Tanaka, H.; Yokoyama, M.; Matsuoka, Y.; Jinzaki, M.; et al. Apparent Diffusion Coefficient Map-Based Texture Analysis for the Differentiation of Chromophobe Renal Cell Carcinoma from Renal Oncocytoma. *Diagnostics* **2022**, *12*, 817. [CrossRef]
59. Chieregato, M.; Frangiamore, F.; Morassi, M.; Baresi, C.; Nici, S.; Bassetti, C.; Bnà, C.; Galelli, M. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. *Sci. Rep.* **2022**, *12*, 4329. [CrossRef]
60. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
61. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412–D419. [CrossRef]
62. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
63. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
64. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [CrossRef] [PubMed]
65. Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8700–8704. [CrossRef] [PubMed]
66. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S.H. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins* **1999**, *35*, 401–407. [CrossRef]
67. Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341. [CrossRef]
68. Chou, K.-C. Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477–483. [CrossRef]
69. Wang, J.; Li, J.; Yang, B.; Xie, R.; Marquez-Lago, T.T.; Leier, A.; Hayashida, M.; Akutsu, T.; Zhang, Y.; Chou, K.-C.; et al. Bastion3: A two-layer ensemble predictor of type III secreted effectors. *Bioinformatics* **2019**, *35*, 2017–2028. [CrossRef]
70. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
71. Acharjee, A.; Larkman, J.; Xu, Y.; Cardoso, V.R.; Gkoutos, G.V. A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Med Genom.* **2020**, *13*, 178. [CrossRef]
72. Chen, R.C.; Dewi, C.; Huang, S.W.; Caraka, R.E. Selecting critical features for data classification based on machine learning methods. *J. Big Data* **2022**, 7. [CrossRef]
73. Yang, Z.; Jin, M.; Zhang, Z.; Lu, J.; Hao, K. Classification Based on Feature Extraction For Hepatocellular Carcinoma Diagnosis Using High-throughput Dna Methylation Sequencing Data. *Procedia Comput. Sci.* **2017**, *107*, 412–417. [CrossRef]
74. Honaas, L.; Hargarten, H.; Hadish, J.; Ficklin, S.P.; Serra, S.; Musacchi, S.; Wafula, E.; Mattheis, J.; dePamphilis, C.W.; Rudell, D. Transcriptomics of Differential Ripening in 'd'Anjou' Pear (Pyrus communis L.). *Front. Plant Sci.* **2021**, *12*, 609684. [CrossRef] [PubMed]