Modeling Primer-Template Interactions using BERT Tokenizer to Predict PCR Amplification with Attention-BiLSTM

Niloofar Latifian[a,*], Naghmeh Nazer[b], Amir Masoud Jafarpisheh[b], Babak Hossein Khalaj[b]

[a] *Department of Computer Engineering, Sharif University of Technology, Tehran, Iran*
[b] *Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran*

Abstract

Polymerase Chain Reaction (PCR) is a widely used molecular biology technique to amplify DNA sequences. PCR amplification is affected by factors such as binding dynamics and primer-template interactions. This study aims to reduce the time and cost of the experiment by predicting PCR outcomes based on these factors. To achieve this, we first identify the most stable binding sites for each primer-template pair by calculating the Gibbs free energy. Then, we propose a unique labelling strategy that captures primer-template interactions in the binding sites by analyzing match and mismatch positions. We categorize a set of English words into two semantically distinct groups: one for match positions and another for mismatch positions. Words within each group have a higher cosine similarity to one another than to words in the opposing group. We assign the corresponding word to each base pair based on whether it is a match or a mismatch. The labelled sequence is then tokenized with BERT, serving as input to an attention Bi-LSTM model. Achieving 96.3% accuracy, this approach significantly outperforms prior methods and pioneers BERT-based analysis in primer-template bindings.

Keywords: PCR amplification, Primer-template interaction, BERT tokenizer, Attention Bi-LSTM

*Corresponding author.

Email addresses: `niloofar.latifian@gmail.com` (Niloofar Latifian), `naghme93@gmail.com` (Naghmeh Nazer), `jafarpisheh@ee.sharif.edu` (Amir Masoud Jafarpisheh), `khalaj@sharif.edu` (Babak Hossein Khalaj)

1. Introduction

Polymerase Chain Reaction (PCR) is a widely utilized molecular biology technique that enables the exponential amplification of specific DNA sequences. It allows researchers to generate millions or even billions of copies of a target DNA fragment from a minimal starting sample [1]. The amplification of a specific segment of DNA enables its analysis in various studies. One of the recently popular and well-known applications of this process is the detection of the coronavirus in patients [2]. The PCR process begins with extracting a DNA template containing the target sequence, which can be sourced from cells, tissues, or environmental samples. Once the template is prepared, short DNA sequences called primers are designed and synthesized to bind to regions flanking the target sequence. The primers that bind to the antisense strand (3' to 5') and sense strand (5' to 3') are named forward primer and reverse primer respectively. These primers act as initiation sites for DNA synthesis. The design of specific and efficient primers is crucial to ensuring the successful amplification of the target sequence, as they must precisely complement both ends of the DNA fragment to be amplified [3].

In a laboratory, several factors critically impact the success of PCR amplification. Parameters such as primer design, template DNA concentration, annealing temperature, and cycle conditions must be carefully optimized for each experiment. Important considerations in primer design include Gibbs free energy (ΔG) for stable primer-template binding [4], avoiding 3' mismatches to ensure proper extension [5], [6], and maintaining a balanced GC content (ideally 40-60%) [7] to promote stable hybridization. Additionally, secondary structures like hairpins or dimers must be minimized to ensure proper binding [8].

Various software tools have been developed for primer design like Primer3 [9], OLIGO7 [10], and PrimerSelect [11]. These tools generate suitable primers by analyzing the template sequence and considering various user-input parameters. Due to the involvement of numerous parameters simultaneously, selecting appropriate primers is complicated and time-consuming. Some prior works tried to optimize the PCR amplification process by selecting better primers [12, 13, 8] or improving its efficiency [14, 15]. In [12, 13], a teaching-learning-based method for selecting the best primer to bind to the template was proposed. This approach employs iterative algorithms that

mimic teaching and learning dynamics to refine primer selection based on multiple criteria, potentially leading to more effective and specific amplification. In [8], An algorithm was developed to prioritize the similarity between designed primers and their target sequences. By focusing on this crucial aspect, the algorithm aims to enhance specificity and reduce off-target amplification. In [14], an Artificial Neural Network was designed with nine experimental inputs of PCR experiment to estimate its efficiency [14]. This approach allows for more accurate predictions of amplification success based on input conditions. In [15], a web tool was designed to identify some parameters that affect PCR efficiency using a statistical approach. This platform enables researchers to systematically evaluate the impact of various factors on amplification outcomes.

There are some parameters that may not be prioritized in the software's design. As a result, designed primers may fail to bind correctly to the target template during the experiment, leading to unsuccessful amplification. Additionally, in certain cases, researchers may aim to determine whether each given primer-template pair will successfully amplify, based on an existing dataset of templates and primers. In these cases, the ability to predict PCR outcomes by precisely accounting for the parameters could greatly reduce the time and resources spent on the PCR, which is a significant challenge [16]. Furthermore, improving the prediction of primer binding specificity and minimizing off-target amplification would decrease the likelihood of false positives and negatives, enhancing the overall reliability of the experiment.

Despite the extensive application of PCR in molecular biology, relatively few studies have focused specifically on predicting the outcome of PCR amplification. Cordaro et al utilized a random forest classifier to Predict PCR results for a diverse set of templates ranging from human, plant, monkey, bacteria, and plasmid DNA [16]. This study included 290 PCR reactions consisting of 109 unique primer pairs. They considered 15 different features of the experiment to train their model, which yielded 81% accuracy. Kronenberger et al also used a random forest classifier to predict PCR amplification results based on 10 parameters affecting the experiment [17]. The study focused on environmental DNA (eDNA) and paired 10 qPCR assays with 82 synthetic gene fragments. They conducted 530 specificity tests, achieving accuracies of 92.4% with SYBR Green and 96.5% with TaqMan hydrolysis probes. Döring et al presented a dataset of 47 immunoglobulin heavy chain variable sequences and 20 primers [18]. They developed a logistic regression model considering 22 parameters for predicting the amplification status

with an AUC of 95%. Kayama et al expressed the relationship between the primer sequence and the template as a five-lettered word and then predicted the PCR results using an RNN model [19]. They reached 80% accuracy on a dataset of 72 PCR primers capable of amplifying 31 DNA templates. Bai et al predicted PCR amplification results using the same dataset and same modeling approach, with an attention Bi-LSTM gaining 82% accuracy [20].

While numerous tools exist for primer design, predictive models that consider various parameters influencing PCR efficiency—such as primer-template interaction, reaction conditions, and the potential for non-specific binding—are limited. Furthermore, the existing methods do not have high accuracy, and there is still a need for improvement. On the other hand, some of these models require examining various parameters involved in the experiments and lack comprehensiveness. This gap highlights the need for more systematic approaches to predict PCR results, especially as PCR continues to be a fundamental technique in genetic and genomic studies.

In this study, we aim to predict the outcome of PCR amplification more accurately than previous methods. For this purpose, first, we use a feature selection method to select the most important features in predicting the outcome. Then, we utilize a BERT tokenizer to model each match and mismatch in primer-template binding. After all, by developing a Bidirectional Long Short Term Memory model with an attention mechanism (Attention Bi-LSTM) [21] we predict the amplification result with 96.3% accuracy. To the best of our knowledge, this is the first study to use a BERT tokenizer [22] for modeling interactions between base pairs. Our method aims to simplify experimental workflows and improve the efficiency of PCR-based applications.

2. Methods

2.1. Datasets and Data Prepration

To predict the outcomes of PCR, it is essential to first collect appropriate data. This involves conducting and recording the results of the PCR process on a set of primers and DNA samples, categorizing the outcomes as either "amplified" or "not amplified." There are few datasets in the scientific literature that document experimental features and their final results in this manner, particularly those that include "not amplified" samples. In this study, we utilized datasets provided by Döring et al [18], and Kayama et al [19] that contain both desired outcomes.

In the Kayama et al dataset, 72 pairs of PCR primers with lengths ranging from 19 to 22 bp, were designed with Primer3 software to amplify 31 double-stranded templates, with lengths ranging from 435 to 481 bases. Templates were synthetically generated using the OE-PCR technique for partial 16S rRNA nucleotide sequences (v6-v8). The design of primers was based on their specificity for the corresponding templates, targeting an amplification length of approximately 100-150 bp. It is assumed that primer-template binding is not limited to the primer's full length and only a part of 3' may bind.

It is important to note that the initial dataset is imbalanced, with approximately 75% of the data being unamplified and about 25% amplified. To improve and comprehensively predict amplification outcomes, it is necessary to balance the dataset before model implementation, ensuring an equal number of samples for each category. Consequently, we oversampled the minority class in our dataset by repeating the samples. As a result the dataset containing 2,232 entries, was expanded to 3,362 entries comprising 1,681 amplified samples and 1,681 unamplified samples.

In contrast to the Döring et al dataset, the Kayama et al dataset does not provide direct documentation of various experimental features. The data available in their study includes only the sequences of the template and primers, along with the corresponding amplification results. Nevertheless, features such as Gibbs free energy and mismatch information can be analyzed solely based on the primer and template sequences.

2.2. Feature Selection

Given the diversity of features across the dataset, it was crucial to identify the features that had the greatest impact on PCR amplification outcomes. To achieve this goal, we applied a feature selection method to the Döring et al dataset. This dataset includes data on 22 different features related to PCR and the final amplification results. Due to the common nature of the problem, we can select the most important features and then analyze these features in the dataset of Kayama et al to perform prediction using our proposed model. We used an embedded method to identify features that significantly influenced PCR results. In this method, the subset of selected features achieves the best accuracy when derived during the construction of a classifier [23]. Embedded methods leverage each iteration of the model training process for feature selection, effectively identifying features that provide the greatest benefit for enhancing data training. During the training phase,

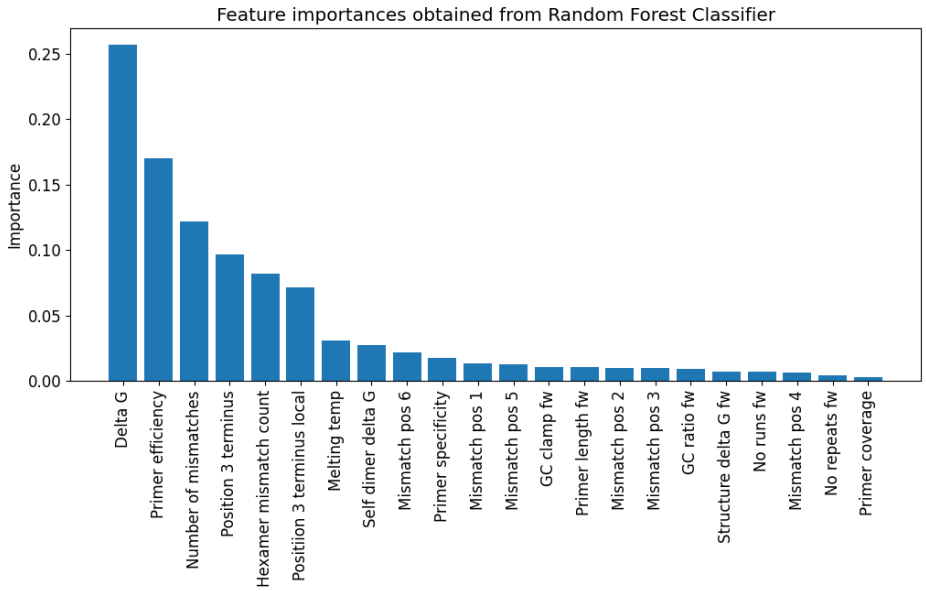


Figure 1: Results of feature selection using Random Forest Classifier on the first dataset. We chose the first six features as the most important features for predicting PCR amplification results.

the classifier modifies its internal parameters and assesses the appropriate weights for each feature in order to achieve the best classification accuracy [24]. This can be achieved using various classifiers, such as the random forest classifier used in our study. By applying this algorithm to the dataset, we identified several key features that are most important for predicting PCR amplification, shown in Figure 1. While Gibbs free energy is an essential feature that specifies the success of binding, mismatches between the primer and template at different binding sites are also crucial. The number and location of mismatches directly influence the stability of the binding, and thus the amplification outcome.

2.3. Binding site identification

Primers can bind to various positions on the template, and the stability of these bindings is determined by the Gibbs free energy (ΔG). The lower the ΔG, the more stable the binding. Because the complementary region is short enough, the most stable pairing is likely to be the priming position. To identify the most stable binding position for each primer-template pair, we should slide the primer on the template and calculate ΔG for each position

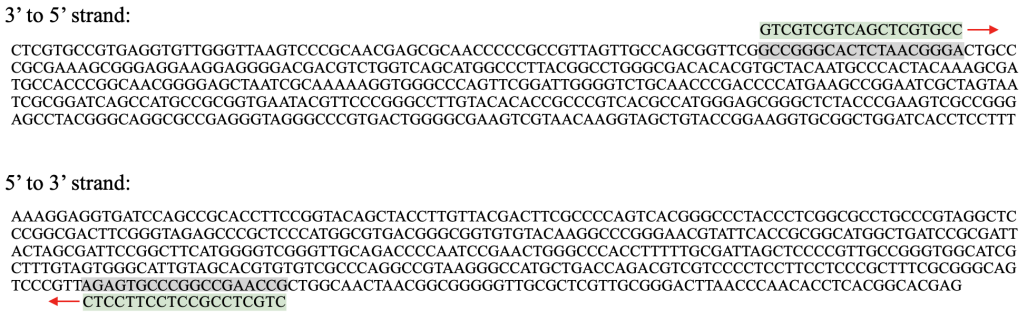


Figure 2: Sliding forward and reverse primer on the corresponding template strand and calculating the ΔG of each place. The place with minimum ΔG may be the binding position.

of binding and choose the place with minimum ΔG, as shown in Figure 2. ΔG is defined as below,

$$\Delta G = \Delta H - T\Delta S. \tag{1}$$

In this formula, ΔH and ΔS are enthalpy and entropy of sequence, respectively. T is the annealing temperature which is 56° for this experiment. We can calculate ΔH and ΔS using the approach presented in the study of Horne et al [25]. In this approach for a pair of primer-template, ΔH is calculated by adding up the enthalpy of each pair of neighbouring bases as shown below,

$$\begin{aligned}
&\Delta H \begin{pmatrix} GTCGTCGTCAGCTCGTGCC \\ GCCGGGCACTCTAACGGGA \end{pmatrix} \\
&= \Delta H \begin{pmatrix} GT \\ GC \end{pmatrix} + \Delta H \begin{pmatrix} TC \\ CC \end{pmatrix} + \Delta H \begin{pmatrix} CG \\ CG \end{pmatrix} \\
&+ \ldots + \Delta H \begin{pmatrix} GC \\ GG \end{pmatrix} + \Delta H \begin{pmatrix} CC \\ GA \end{pmatrix}.
\end{aligned} \tag{2}$$

ΔS can be calculated using the same way. ΔH and ΔS for each type of neighbouring base pairs are proposed in the same study. We calculate ΔG using this method for both forward and reverse primers. After this calculation for all of the possible binding positions for each primer, we choose places for binding positions of forward and reverse primers with minimum ΔG that are at least 100 bp apart.

With the identification of the binding site which was the most significant feature for determining the output, we can now focus on other features. As

shown in Figure 1, one of the most important features is the primer efficiency. Since the primers were previously designed using Primer3, this feature has already been evaluated and does not require further consideration at this stage. Other important features in determining the outcome are based on match and mismatch information, such as their number and location.

2.4. Labelling match and mismatches positions

At this point, the problem can be defined as working with sequences composed of the nucleotides A, T, C, and G. In the regions where two sequences bind, we are interested in identifying matches and mismatches. It was noted that the primer-template binding site can be smaller than the length of the primer. Therefore, matches or mismatches may also occur in nucleotides before or after the main binding site. Thus, we need to display each binding position in a way that distinguishes the matched regions from mismatched ones at the binding site and before and after it. To achieve this goal, our idea is to use two groups of English words for 1) match places and 2) mismatch places with a similar meaning to "match" and "mismatch" words. We need meaningful similar words because after selecting these words, we will tokenize them with a BERT tokenizer to use them as input for our attention Bi-LSTM model.

By measuring the cosine similarities, we found the words "match", "obtained", "equal" and "produced" suitable for match places, and the words "mismatch", "discrepancy", "disparity" and "misfit" suitable for mismatch places in the binding site and non-binding site positions of forward and reverse primers. After these assignments, each primer-template pair will be modeled as a sequence containing these words. An example of the labelling of forward and reverse primer binding sites and the resulting sequence from this process is illustrated in Figure 3. For each base pair at each position in the forward primer's binding to the template, we assign the corresponding word. This process is repeated for the reverse primer. Green words indicate the binding site of the primer to the template and black words indicate before and after the binding site. Two sequences are generated for both forward and reverse primers. These two sequences are then concatenated to produce the final sequence. This process is repeated for each primer-template pair to generate the final sequence for all combinations. This labelling approach provides a comprehensive representation of the primer-template interactions, helping to systematically analyze primer binding specificity across different pairs.

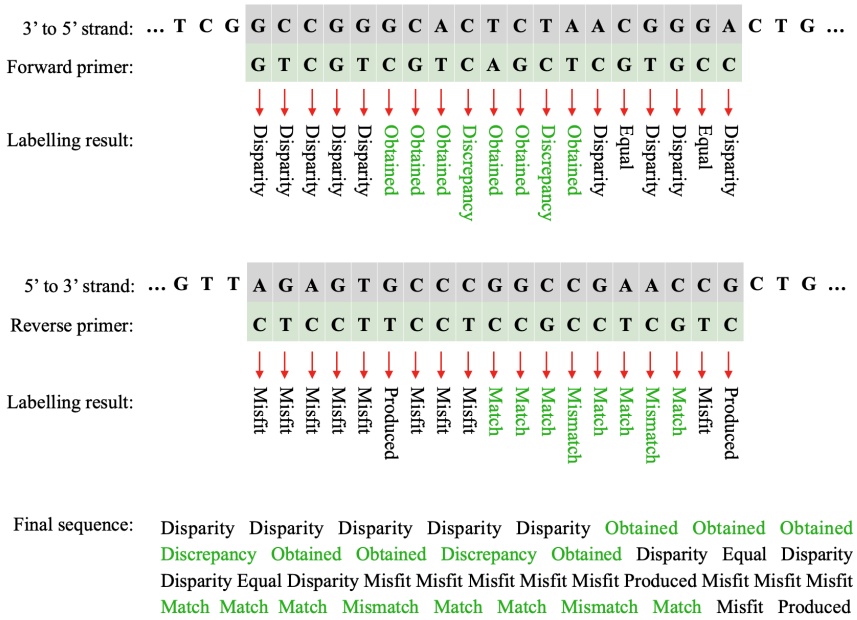


Figure 3: Primer-template labelling based on matches and mismatches in each position. We assign words to each base pair based on its class (match or mismatch), type (reverse or forward), and position (at the binding site or before and after it. The final sequence is generated by concatenating the labelling results for forward and reverse primers.

2.5. Data Augmentation

Since the dataset is relatively small, we employed a data augmentation technique to enhance the generalizability of our model [26]. By systematically generating variations of the existing dataset, we increased the diversity of training samples and enabled the model to better adapt to unseen data. We doubled the dataset and replaced "match" with "pass", "mismatch" with "dissidence", "obtained" with "align", and "discrepancy" with "dissimilation" in new sequences. The Tabel 1 represents words assigned to each type of position in the augmented dataset. For two categories of matches and mismatches, the words of each group are more similar to each other than to the words of the other group.

2.6. BERT Tokenizer

We used a BERT tokenizer to tokenize each sequence by its words. BERT tokenization transforms text into tokens that represent individual words or

Primer		Corresponding label	
		At the binding site	Before/after the binding site
Forward	Match	match pass	equal
	Mismatch	mismatch dissidence	disparity
Reverse	Match	obtained aligned	produced
	Mismatch	discrepancy dissimilation	misfit

Table 1: Words assigned to each type of primer-template bindings in the augmented dataset

subwords, allowing the model to understand and process the text effectively [22]. This tokenization captures contextual relationships and semantic meanings, which are crucial for tasks such as text classification or sequence prediction in models like Bi-LSTM. We use this tokenization to capture patterns of match and mismatches in the primer-template pairs of PCR.

3. Results

In this study, we utilized a Bi-LSTM model with attention mechanisms to predict PCR amplification results. Bi-LSTM models are a type of Recurrent Neural Network (RNN) commonly employed in Natural Language Processing (NLP) tasks. The bidirectional nature of a Bi-LSTM allows it to process sequences from both ends. Like standard LSTMs, these models can remember long-term dependencies in data. We add an attention mechanism that considers the importance of different parts of the input sequence dynamically, to concentrate on more essential information and improve context understanding. To guarantee the accuracy of model prediction, we used a 5-fold cross-validation for 250 epochs. Furthermore, we did a train-test split with a ratio of 80/20 for performance evaluation. Adam optimizer was utilized with a start learning rate of 0.0004 and a scheduler to reduce the learning rate when the validation accuracy stopped improving.

Accuracy, sensitivity, and specificity are employed to assess the proposed model, each reflects the strengths of the algorithm from a particular per-

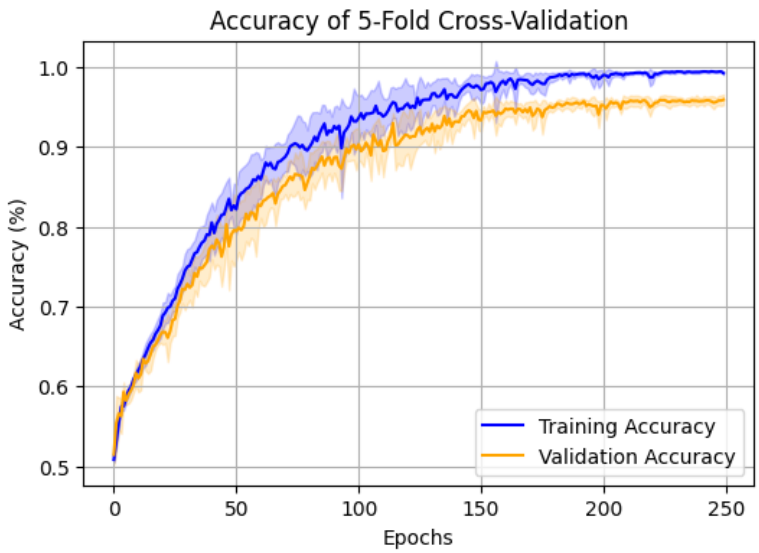


Figure 4: Training and Validation accuracy results for 5-fold cross-validation.

spective. We have a sensitivity of 0.980 ± 0.013 for validation data and a sensitivity of 0.977 ± 0.005 for our test data. The specificity reaches 0.931 ± 0.008 and 0.948±0.010 for validation and test data, respectively. Our developed model achieved a validation accuracy of 0.951±0.005% (averaged over the last 100 epochs) and a test accuracy of 96.3±0.009%. Figure 4 illustrates the training and validation accuracy results for 5-fold cross-validation.

4. Discussion

In the study of Kayama et al, Bai et al, and ours, the same dataset is utilized. In the Kayama et al and Bai et al studies, the coded sequence is composed of 25 English characters in lowercase and uppercase forms, based on the position and type of nucleotide pairs. The characters used correspond to each type of connection between nucleotides, with every five consecutive characters forming a word. Some of these words are repeated due to their significance. The final accuracy of the models used by Kayama et al and Bai et al reached 80% and 82% accuracy, respectively.

In our proposed model, We presented a new labelling strategy based on the location and number of mismatches in the forward and reverse primer-template binding. We claim that when the aim is to identify the binding

pattern of the primer to the template, the specific types of nucleotides involved are not critical. What matters is whether the paired nucleotides are matches or mismatches. For instance, there is no distinction between C-G and A-T connections, as both are considered matches. Similarly, T-T and A-A connections are both mismatches and in our proposed model, there is no need to assign separate codes for each. The type of bases present in a nucleotide connection only affects the calculation of Gibbs free energy since each consecutive pair of nucleotides has its own enthalpy and entropy values. In our proposed model, the calculation of Gibbs free energy occurs before word formation. After selecting the appropriate location for primer binding, the specific nucleotide types are no longer necessary. It suffices to know whether the connection is a match or a mismatch. Moreover, to emphasize features like closeness to the 3' terminal for the RNN model, their models repeat words based on their positional placement. In contrast, our proposed model does not require such repetition of words. This is because it labels connections between primers and templates based solely on matches and mismatches. As a result, identifying all relevant features becomes more straightforward, especially when we utilize a BERT tokenizer to recognize patterns in the sequence considering their type and position. The words we chose for each of the match and mismatch groups are more similar to each other than the other group based on their cosine similarity. This leads to similar meanings when we tokenize these words in the sequence. As shown in Table 2 the accuracy in our proposed model is significantly higher with about a 13% increase for validation data and a 9% increase for test data. The significant difference in accuracy between our model (96.3%) and that reported in previous studies (80% and 82%) indicates that concentrating on match-related features leads to improved prediction accuracy.

Additionally, our model shows improved specificity and, in particular, significantly higher sensitivity compared to previous models, providing further evidence of its superior performance. The notably increased sensitivity indicates a considerable improvement in detecting true positives over prior models. As the original dataset is imbalanced, the models proposed in the prior studies fail to detect the majority of actual positive cases, likely misclassifying them as negatives. Kayama et al evaluated their model on a downsampled version of the dataset, reaching a sensitivity of 75.1% and 47.1% in validation and test, respectively. Although sensitivity increases with downsampling in their model, accuracy and specificity tend to decrease. By reducing the number of majority class (negative) samples to balance with the minority

Reference	Method	Validation			Test		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Kayama et al [19] whole data	LSTM	0.563 ± 0.062	0.882 ± 0.024	0.803 ± 0.029	0.114 ± 0.092	0.899 ± 0.022	0.862 ± 0.020
Kayama et al [19] downsampled data	LSTM	0.751 ± 0.031	0.686 ± 0.021	0.718 ± 0.022	0.471 ± 0.092	0.661 ± 0.056	0.652 ± 0.052
Döring et al [18]	Attention-BiLSTM	0.570±0.06	0.905±0.03	0.822±0.02	0.117±0.09	0.914±0.03	0.874±0.02
Current study	Attention-BiLSTM	0.980 ± 0.013	0.931 ± 0.008	0.951±0.005	0.977 ± 0.005	0.948±0.010	0.963±0.009

Table 2: Comparing results with prior studies

class (positive), the model becomes more likely to misclassify majority class samples as positive, resulting in lower accuracy and specificity.

In addition to comparing our results with studies that used the same dataset, we reviewed other relevant works that relied on a predefined table of features, predicting the outcome with machine learning models. Cordaro et al [16] and Döring et al [18] have achieved accuracies of 81% and 95%. Kronenberger et al yielded 92.4% and 96.5% for two different types of probes in their study. The feature-based approaches require comprehensive documentation of all relevant features, which can be challenging, especially as new data characteristics emerge. In contrast, our method circumvents this limitation by predicting PCR outcomes directly from the template and primer sequences, without the need for extensive feature documentation. This streamlined approach not only simplifies the predictive process but also enhances adaptability to new data, making it more efficient and scalable.

Furthermore, the lack of augmentation in all of the prior works while having small datasets, limits the robustness and generalizability of their models, as small datasets can lead to overfitting. In contrast, our study incorporates augmentation, which improves model performance and reliability, especially in low-data settings.

5. Conclusion

In this study, we predicted PCR amplification using BERT tokenization and an attention Bi-LSTM model. First, we identified the most important features for the amplification result prediction based on a dataset containing different features, which resulted in ΔG, primer efficiency, and some other features related to match and mismatch positions. By analyzing each binding place between each pair of primer and template in the dataset, we found the place with the least ΔG, which is the expected binding site. The primer-

template binding was then labelled based on matches and mismatches at the binding site or immediately before and after it. Different words were assigned to represent matches and mismatches at the binding site and its adjacent positions, resulting in a unique sequence for each primer-template combination. We then used a BERT tokenizer to tokenize each sequence based on its words. The collection of these encoded sequences was treated as input for an attention Bi-LSTM. By running the model on this input data with 5-fold cross-validation, a sensitivity of 0.977 ± 0.005%, a specificity of 0.948±0.010%, and an accuracy of 96.3±0.009%, was achieved. Our presented model not only has remarkably higher accuracy, sensitivity, and specificity but is also simpler than the models presented in previous studies by proposing more meaningful and applicable labelling. This approach enables predictions of amplification results based solely on template and primer sequences, thereby saving time and costs associated with conducting PCR experiments.

6. CRediT authorship contribution statement

Niloofar Latifian: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing. **Naghmeh Nazer:** Conceptualization, Methodology, Project administration, Writing – review and editing. **Amir Masoud Jafarpisheh:** Investigation, Methodology, Writing – review and editing. **Babak Hossein Khalaj:** Project administration, Supervision, Writing – review and editing.

7. Declaration of competing interest

The authors have no conflict of interest to declare.

8. Acknowledgment

We thank Professor Daiji Endoh at the Department of Radiation Biology, School of Veterinary Medicine, Rakuno Gakuen University for kindly providing the data used in this study.

9. Data and Code Availability

The data and code supporting this work are openly available at: https://github.com/niloofarlatifian/PCR-Amplification-Prediction.

10. Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

11. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to improve the manuscript's readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

[1] e. a. Green MR, The basic polymerase chain reaction (pcr), Cold Spring Harb Protoc. (2018).

[2] A. Tahamtan, A. Ardebili, Real-time rt-pcr in covid-19 detection: issues affecting the results, Expert Review of Molecular Diagnostics 20 (5) (2020) 453–454.

[3] e. a. Bustin SA, Parameters for successful pcr primer design, Methods Mol Biol (2020) 5–22.

[4] e. a. Alexander A. Lomzov, Evaluation of the gibbs free energy changes and melting temperatures of dna/dna duplexes using hybridization enthalpy calculated by molecular dynamics simulation, Phys. Chem. B (2015).

[5] e. a. S.Kwok, Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies, Nucleic Acids Research 18 (4) (1990) 999–1005.

[6] M. P. C. e. a. Brittany Knight, Taylor Otwell, Machine learning models to predict the impact of template mismatches on polymerase chain reaction (pcr) assay performance., MPREPRINT (Version 1) available at Research Square (2024).

[7] G. S. D. C W Dieffenbach, T M Lowe, General concepts for pcr primer design, PCR Methods Appl (1993).

[8] D. Guha Roy, A new algorithm for primer design, Electronic Thesis and Dissertation Repository, 2019.

[9] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, S. G. Rozen, Primer3—new capabilities and interfaces, Nucleic Acids Research 40 (15) (2012) e115–e115.

[10] W. Rychlik, Oligo 7 primer analysis software, PCR primer design (2007) 35–59.

[11] K. J. Graham, M. J. Holland., Primerselect: a transcriptome-wide oligonucleotide primer pair design program for kinetic rt-pcr-based transcript profiling., Methods in enzymology (2005) 35–59.

[12] C. YH, Computational intelligence-based polymerase chain reaction primer selection based on a novel teaching-learning-based optimisation, IET Nanobiotechnol (2014).

[13] Y.-H. Cheng, Estimation of teaching-learning-based optimization primer design using regression analysis for different melting temperature calculations, IEEE Transactions on NanoBioscience 14 (1) (2015) 3–12. doi:10.1109/TNB.2014.2352351.

[14] D. X. Y. Q. e. a. Li, Y., Development and validation of a new pcr optimization method by combining experimental design and artificial neural network., Appl Biochem Biotechnol (2010).

[15] W. J. . E.-C. M. Mallona, I., pcrefficiency: a web tool for pcr amplification efficiency prediction., BMC Bioinformatics (2011). doi:10.1186/1471-2105-12-404.

[16] e. a. Nicholas J. Cordaro, Optimizing polymerase chain reaction (pcr) using machine learning, bioRxiv (2021).

[17] e. a. John A. Kronenberger, ednassay: A machine learning tool that accurately predicts qpcr cross-amplification, Molecular Ecology Resources (2022).

[18] e. a. Matthias Döring, Modeling the amplification of immunoglobulins through machine learning on sequence-specific features, Scientific RepoRts (2019).

[19] e. a. Kotetsu Kayama, Prediction of pcr amplification from primer and template sequences using recurrent neural network, Scientific Reports 11 (2021).

[20] B. Bai, J. Long, Z. Yang, J. Li, P. Wei, An optimization method of primer design based on attention-bilstm, 2023 2nd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (2023) 263–268doi:10.1109/RAIIC59453.2023.10280917.

[21] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), 2016, pp. 207–212.

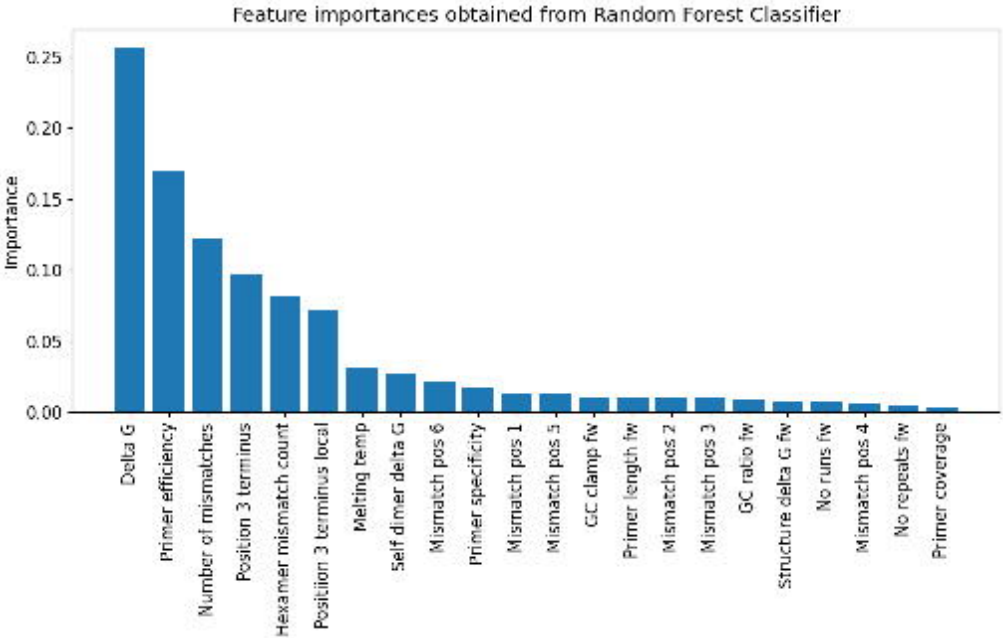
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[23] U. M. Khaire, R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University - Computer and Information Sciences 34 (4) (2022) 1060–1073. doi:https://doi.org/10.1016/j.jksuci.2019.06.012.
URL https://www.sciencedirect.com/science/article/pii/S1319157819304379

[24] K.-L. A. O. J. Pudjihartono N, Fadason T, A review of feature selection methods for machine learning-based disease risk prediction, Front Bioinform (2022).

[25] M. Horne, D. Fish, A. Benight, Statistical thermodynamics and kinetics of dna multiplex hybridization reactions, Biophysical Journal 91 (11) (2006) 4133–4153. doi:https://doi.org/10.1529/biophysj.106.090662.

[26] K. L. P. W. Y. F. C.-T. L. C. C. A. J. P. Y. Z. Zaitian Wang, Pengfei Wang, A comprehensive survey on data augmentation, arXiv (2024).


Feature importances obtained from Random Forest Classifier
Importance
0.25
0.20
0.15
0.10
0.05
0.00
Delta G
Primer efficiency
Number of mismatches
Position 3 terminus
Hexamer mismatch count
Position 3 terminus local
Melting temp
Self dimer delta G
Mismatch pos 6
Primer specificity
Mismatch pos 1
Mismatch pos 5
GC clamp fw
Primer length fw
Mismatch pos 2
Mismatch pos 3
GC ratio fw
Structure delta G fw
No runs fw
Mismatch pos 4
No repeats fw
Primer coverage

3' to 5' strand:

GTCGTCGTCAGCTCGTGCC →

CTCGTGCCGTGAGGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCCGCCGTTAGTTGCCAGCGGTTCGGCCGGGCACTCTAACGGGACTGCC
CGCGAAAGCGGGAGGAAGGAGGGGACGACGTCTGGTCAGCATGGCCCTTACGGCCTGGGCGACACACGTGCTACAATGCCCACTACAAAGCGA
TGCCACCCGGCAACGGGGAGCTAATCGCAAAAAGGTGGGCCCAGTTCGGATTGGGGTCTGCAACCCGACCCCATGAAGCCGGAATCGCTAGTAA
TCGCGGATCAGCCATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACGCCATGGGAGCGGGCTCTACCCGAAGTCGCCGGG
AGCCTACGGGCAGGCGCCGAGGGTAGGGCCCGTGACTGGGGCGAAGTCGTAACAAGGTAGCTGTACCGGAAGGTGCGGCTGGATCACCTCCTTT

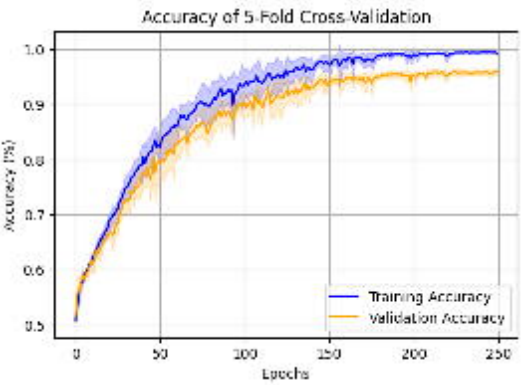
5' to 3' strand:

AAAGGAGGTGATCCAGCCGCACCTTCCGGTACAGCTACCTTGTTACGACTTCGCCCCAGTCACGGGCCCTACCCTCGGCGCCTGCCCGTAGGCTC
CCGGCGACTTCGGGTAGAGCCCGCTCCCATGGCGTGACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCGGCATGGCTGATCCGCGATT
ACTAGCGATTCCGGCTTCATGGGGTCGGGTTGCAGACCCCAATCCGAACTGGGCCCACCTTTTTGCGATTAGCTCCCCGTTGCCGGGTGGCATCG
CTTTGTAGTGGGCATTGTAGCACGTGTGTCGCCCAGGCCGTAAGGGCCATGCTGACCAGACGTCGTCCCCTCCTTCCTCCCGCTTTCGCGGGCAG
TCCCGTTAGAGTGCCCGGCCGAACCGCTGGCAACTAACGGCGGGGGTTGCGCTCGTTGCGGGACTTAACCCAACACCTCACGGCACGAG
← CTCCTTCCTCCGCCTCGTC

3' to 5' strand:	…	T	C	G	G	C	C	G	G	G	C	A	C	T	C	T	A	A	C	G	G	G	A	C	T	G	…
Forward primer:					G	T	C	G	T	C	G	T	C	A	G	C	T	C	G	T	G	C	C				
Labelling result:					Disparity	Disparity	Disparity	Disparity	Disparity	Obtained	Obtained	Obtained	Discrepancy	Obtained	Obtained	Discrepancy	Obtained	Disparity	Equal	Disparity	Disparity	Equal	Disparity				

5' to 3' strand:	…	G	T	T	A	G	A	G	T	G	C	C	C	G	G	C	C	G	A	A	C	C	G	C	T	G	…
Reverse primer:					C	T	C	C	T	T	C	C	T	C	C	G	C	C	T	C	G	T	C				
Labelling result:					Misfit	Misfit	Misfit	Misfit	Misfit	Produced	Misfit	Misfit	Misfit	Match	Match	Match	Mismatch	Match	Match	Mismatch	Match	Misfit	Produced				

Final sequence: Disparity Disparity Disparity Disparity Disparity Obtained Obtained Obtained Discrepancy Obtained Obtained Discrepancy Obtained Disparity Equal Disparity Disparity Equal Disparity Misfit Misfit Misfit Misfit Misfit Produced Misfit Misfit Misfit Match Match Match Mismatch Match Match Mismatch Match Misfit Produced


Accuracy of 5-Fold Cross-Validation
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy (%)
0
50
100
150
200
250
Epochs
Training Accuracy
Validation Accuracy