Hindawi
Scientific Programming
Volume 2021, Article ID 5529202, 10 pages
https://doi.org/10.1155/2021/5529202

Research Article

Deep Neural Network for Somatic Mutation Classification

Haifeng Wang,[1] **Chengche Wang**,[2] **and Hongchun Qu**[1]

[1]*School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China*
[2]*Bank of China Zhejiang Branch, Hangzhou, ZheJiang 310003, China*

Correspondence should be addressed to Chengche Wang; 283226494@qq.com and Hongchun Qu; hcchyu@gmail.com

Received 27 January 2021; Revised 16 February 2021; Accepted 11 March 2021; Published 20 March 2021

Academic Editor: Wenzheng Bao

The detection and characterization of somatic mutations have become the important means to analyze the occurrence and development of cancer and, ultimately, will help to select effective and precise treatment for specific cancer patients. It is very difficult to detect somatic mutations accurately from the massive sequencing data. In this paper, a forest-graph-embedded deep feed-forward network (forgeNet) is utilized to detect somatic mutations from the sequencing data. In forgeNet, the random forest (RF) or Gradient Boosting Machine (GBM) and graph-embedded deep feed-forward network (GEDFN) are utilized to extract features and implement classification, respectively. Three real somatic mutation datasets collected from 48 triple-negative breast cancers are utilized to test the somatic mutation detection performances of forgeNet. The detection results show that forgeNet could make the 0.05%–0.424% improvements in terms of area under the curve (AUC) compared with support vector machines and random forest.

1. Introduction

With the rapid development of new sequencing technology, lots of sequencing omics data have been generated which are processed and analyzed in order to solve biological problems [1–5]. The species without a reference sequence are sequenced again at the genomic level, and the reference sequence of this species can be obtained, which will lay the foundation for the follow-up research and molecular breeding [6–8]. For the species with a reference sequence, whole genome sequencing could detect the mutation sites related to important characters of an organism, including single-nucleotide polymorphism (SNP) and insertion-deletion (InDel), which are molecular basis of individual differences and play an important role in research and industry [9–12].

Somatic mutations occur in the normal body cells including SNPs and InDels. Such mutations will not be passed on to the offspring. Somatic mutations are different from germline mutations, which occur in cells becoming gametes (sperm and egg) [13]. Germline mutations could be passed on to the offspring [14, 15]. Somatic mutations do not make genetic changes for the offspring, but these can cause the changes of the genetic structure for some cells. Many researchers have studied the reasons about cancers [16–20]. The abnormal structures or functions of cellular genetic material could be caused by carcinogenic factors. Most of these abnormalities are not inherited from germ cells, but are caused by new gene mutations in somatic cells. The mutated precancerous cells develop into tumors under the action of some tumor-promoting factors [21–23]. Therefore, most of the tumors can be regarded as a kind of somatic genetic disease [24]. The study of cancer-related somatic variation has an important role for the treatment and prevention of cancer.

Nowadays, lots of machine-learning methods have been utilized to solve biomedical problems [25–29]. However, it is very difficult to detect somatic mutations accurately from the massive sequencing data. In recent years, many researchers have been working on solving this problem. Ding et al. investigated performances of four classical classification methods in order to detect a somatic single-nucleotide variant (SNV) [30]. Shiraishi et al. proposed a novel somatic mutation detection algorithm, namely, Bayesian mutation calling, with whole-exome sequencing data. Also, an empirical Bayesian method was presented to detect somatic

mutation and sequencing errors [31]. Koboldt et al. proposed a variant calling tool, namely, VarScan 2, to discriminate germline mutations from somatic mutations from next-generation sequencing (NGS) data [32]. Sahraeian proposed a new somatic identification method based on the convolutional neural network, which could outperform the previous methods [33]. Yang and Chen proposed an ensemble-method-based flexible neural tree model (FNT) and Radial Basis Function (RBF) to improve the accuracy of somatic mutation identification [34]. Dorri et al. proposed the MuClone method to detect somatic mutations with multiple tumor samples, which could classify mutations into biologically meaningful groups [35].

Recently, Kong and Yu presented a novel classifier based on the feature graph and deep neural network, namely, forgeNet. forgeNet was utilized to process RNA-seq data from public databases, and the results proved that this method was valuable for classification and feature selection for biology data [36]. Thus, in this paper, forgeNet is utilized to detect somatic mutations from the sequencing data. In forgeNet, the random forest and graph-embedded deep feed-forward network are utilized. Three real somatic mutation datasets collected from 48 triple negative breast cancers are utilized to test the somatic mutation detection performances of forgeNet.

The rest of the paper is organized as follows. The second section introduces the detailed forgeNet algorithm. The detail identification process of somatic mutation is also given. The third section proposes three experiments on the forgeNet method. The last section provides many conclusions and possible future research.

2. Methods

2.1. forgeNet. The forest-graph-embedded deep feed-forward network (forgeNet) was proposed by Kong in 2020, which is a novel classification method based on the feature extraction algorithm and deep neural network (DNN). This method has been successfully applied to biology data, so forgeNet is utilized to detect somatic mutations. The forgeNet method contains the following two steps [37].

2.1.1. Feature Extractor Part. In this part, random forest and Gradient Boosting Machine (GBM) are utilized to select the proper features according the training dataset. Suppose that a forest has N decision trees. According to the training dataset, the fitting forest could be obtained as $\mathrm{T}(\theta) = \{\Upsilon_1(\theta_1), \Upsilon_2(\theta_2), \ldots, \Upsilon_N(\theta_N)\}$, where θ denotes the parameters of trees. A binary tree could be viewed as a special case of a graph simultaneously, and a set of graphs could be obtained as follows:

$$\begin{aligned} &\Phi = \{G_1(V_1, E_1), \ldots, G_i(V_i, E_i), \ldots, G_N(V_N, E_N)\}, \\ &i = 1, \ldots, N, \end{aligned} \tag{1}$$

where V_i and E_i are sets of vertices and edges in G_i.

The final feature graph G could be obtained by merging all graphs in graph set Φ, which is prepared for the second step of forgeNet.

$$G = \bigcup_{i=1}^{N} G_i. \tag{2}$$

2.1.2. Neural Network Part. In this part, graph-embedded deep feed-forward networks (GEDFNs) are utilized to tackle with classification problems [37]. The structure of the GEDFN is given as follows:

$$\begin{aligned} Z_1 &= \sigma(X(W_{\text{in}} \Theta G) + b_{\text{in}}), \\ &\ldots \\ Z_{k+1} &= \sigma(Z_k W_k + b_k), \\ &\ldots \\ Z_{out} &= \sigma(Z_l W_l + b_l), \\ y &= \text{soft}\max(Z_{\text{out}} W_{\text{out}} + b_{\text{out}}), \end{aligned} \tag{3}$$

where X is the data matrix with the proper features selected from the first step of forgeNet, Θ denotes the Hadamard product, and W_k and b_k are the weights and bias of the k^{th} hidden layer, respectively.

2.2. Somatic Mutation Identification. In order to test the detection performances of forgeNet and identify somatic mutations, a cross-validation method is utilized, which could solve the overfitting problem [38, 39]. By the K-fold cross-validation method, the detection process of somatic mutations with forgeNet is given as follows (Figure 1):

(1) The feature data of somatic mutations are divided into K groups $(S_1, S_2, \ldots, S_K)$, and the numbers of the samples in K groups are generally equal. K is generally greater than or equal to 2.

(2) Each subset is set as a testing set once, and the remaining K-1 subsets are set as a training set. With the divided training and testing sets, the forgeNet method is fitted. Through K runs, K models will be obtained $(M_1, M_2, \ldots, M_K)$. The area under the curve (AUC) of the testing set of these K models is used as the performance index of the classifier.

3. Experiments

In order to investigate the somatic mutation identification performances of forgeNet, three real somatic mutation datasets are utilized, which were collected from 48 triple negative breast cancers by capturing tumour/normal pairs sequenced with the Illumina genome analyzer [30]. The positive and negative samples of datasets are described in Table 1.

Receiver-operating characteristic (ROC) is utilized to measure the performance of the somatic mutation classification model with any dataset, and area under the curve (AUC) is utilized to quantify the ROC curve. The steeper the ROC curve is, the better the performance of classification is.

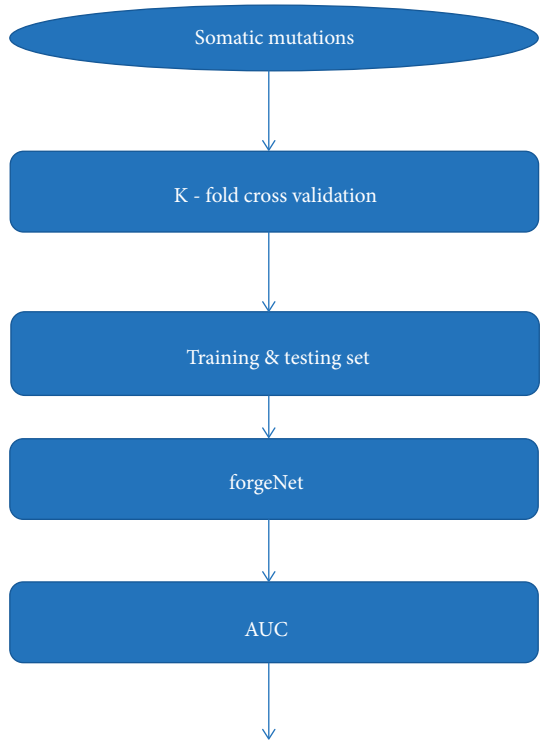

Figure 1: The detection flowchart of somatic mutations with forgeNet.

Table 1: The description of datasets.

	Somatic mutations	Nonsomatic mutations
Dataset 1	775	588
Dataset 2	269	1838
Dataset 3	1015	2354

The value of AUC is between 0.5 and 1. In order to test the detection performances of forgeNet, SN, SP, Acc, MCC, and $F1$ are utilized, which are defined in equation (4). Support vector machines (SVM) [40, 41] and random forest (RF) [42, 43] are also utilized to identify somatic mutations with three real datasets in order to compare the performances of forgeNet.

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{Acc} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \\ F1 &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \end{aligned} \tag{4}$$

The detection results of somatic mutations of SVM, RF, and forgeNet are listed in Table 2 with three datasets. For dataset 1, forgeNet has the highest SN performance, which reveals that forgeNet could identify more true somatic mutations. RF has higher SP than forgeNet and SVM, which shows that RF could identify more true nonsomatic mutations. Overall, $F1$ performances show that RF performs better than forgeNet and SVM, but RF and forgeNet have the extremely close results. For dataset 2, forgeNet and SVM have the same SN performance, which is 0.933. In terms of SP, RF has the best performance, which is 0.997. The Acc result reveals that forgeNet and RF could detect the same numbers of true somatic and nonsomatic mutations. But, in terms of $F1$, forgeNet performs best. For dataset 3, in terms of SN, forgeNet has better performance, while RF has the better SP performance. Overall, forgeNet has higher $F1$ performance than RF.

The identification of AUC performances of three methods (forgeNet, SVM, and RF) by 10-fold cross validation with dataset 1, dataset 2, and dataset 3 is depicted in Figures 2, 3, and 4, respectively. From Figure 2, the ROC curves of RF and forgeNet are very close, which are better

Table 2: The performances of somatic mutation detection by SVM, RF, and forgeNet.

		SN	*SP*	*Acc*	*MCC*	*F*1
Dataset 1	forgeNet	0.986	0.954	0.972	0.943	0.976
	SVM	0.977	0.956	0.968	0.935	0.972
	RF	0.983	0.964	0.975	0.949	0.978
Dataset 2	forgeNet	0.933	0.989	0.981	0.917	0.928
	SVM	0.933	0.967	0.963	0.847	0.865
	RF	0.870	0.997	0.981	0.915	0.923
Dataset 3	forgeNet	0.971	0.986	0.982	0.957	0.97
	SVM	0.974	0.978	0.977	0.946	0.962
	RF	0.961	0.991	0.982	0.956	0.969

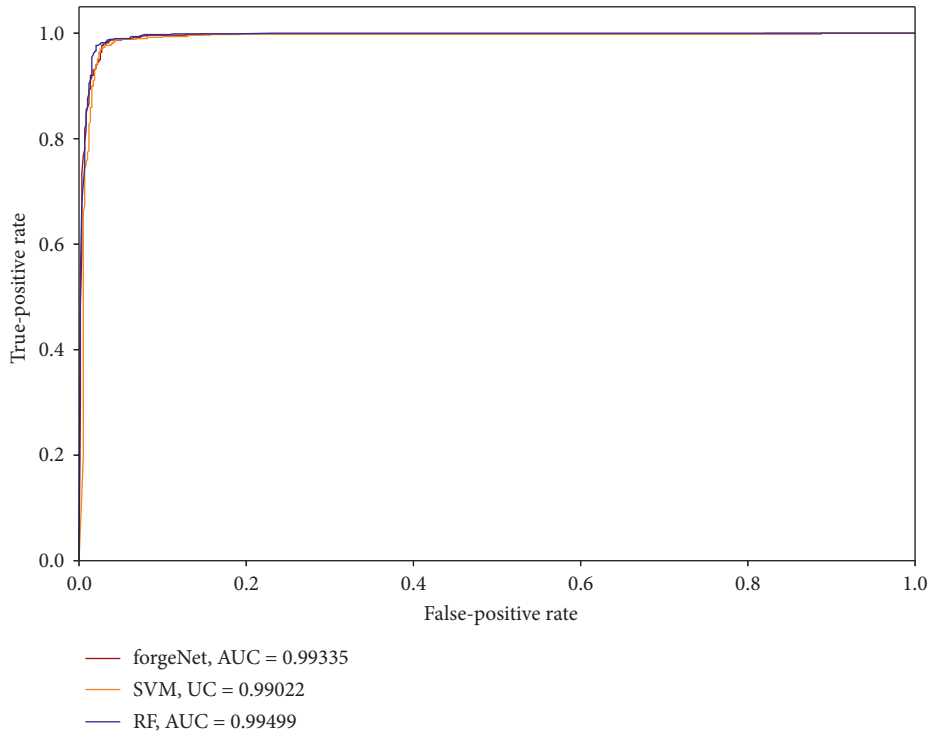

Figure 2: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 1 and 10-fold cross validation.

than that of SVM. RF could obtain the best AUC value, which is 0.99499. forgeNet has the second better AUC value, which is 0.16% lower than that of RF and 0.32% higher than that of SVM. From Figure 3, it could be seen that forgeNet has better ROC curve than RF and SVM with dataset 2. forgeNet could obtain the highest AUC value, which is closer to 1, 0.424% higher than that of SVM and 0.05% higher than that of RF. For Figure 4, with dataset 3, forgeNet and RF have the closer ROC curves, which are better than that of SVM. In terms of AUC value, forgeNet is 0.24% higher than SVM and 0.105% higher than SVM. Through the identification results of three datasets, we can see that forgeNet could obtain better performances than SVM and RF when the ratio of somatic mutations is low.

In order to investigate the performance of forgeNet further, forgeNet, SVM, and RF are utilized to identify somatic mutations with dataset 2 and dataset 3 by 3-fold cross validation, 5-fold cross validation, and 8-fold cross validation, respectively. By 3-fold cross validation, the identification of ROC curves and AUC values of three methods is depicted in Figures 5 and 6 with dataset 2 and dataset 3, respectively. From Figure 5, it could be seen that, in terms of AUC, forgeNet is 0.278% higher than SVM and 0.425% higher than RF. Figure 6 reveals that, in terms of AUC, forgeNet is 0.328% higher than SVM and 0.028% higher than RF.

By 5-fold cross validation, the identification of ROC curves and AUC values of three methods is depicted in Figures 7 and 8 with dataset 2 and dataset 3, respectively. From Figure 7, it could be seen that, in terms of AUC, forgeNet is 0.167% higher than SVM and 0.388% higher than RF. Figure 8 shows that, in terms of AUC, forgeNet is 0.27% higher than SVM and 0.05% higher than RF. By 8-fold cross validation, the identification of ROC curves and AUC values of three methods is depicted in Figures 9 and 10 with dataset 2 and dataset 3, respectively. From Figure 9, it could be seen

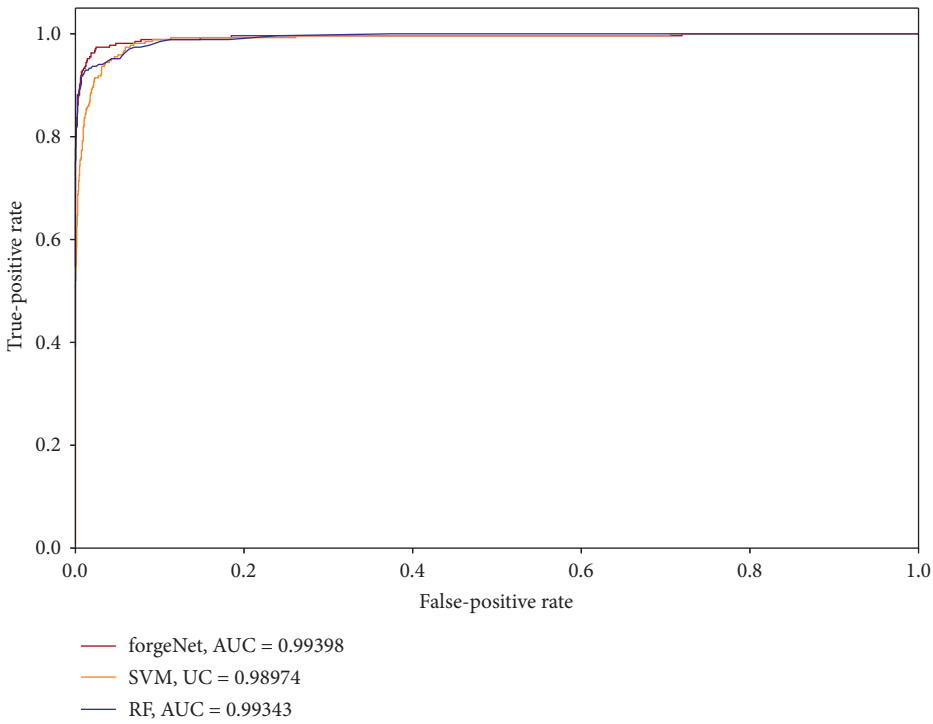

Figure 3: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 2 and 10-fold cross validation.

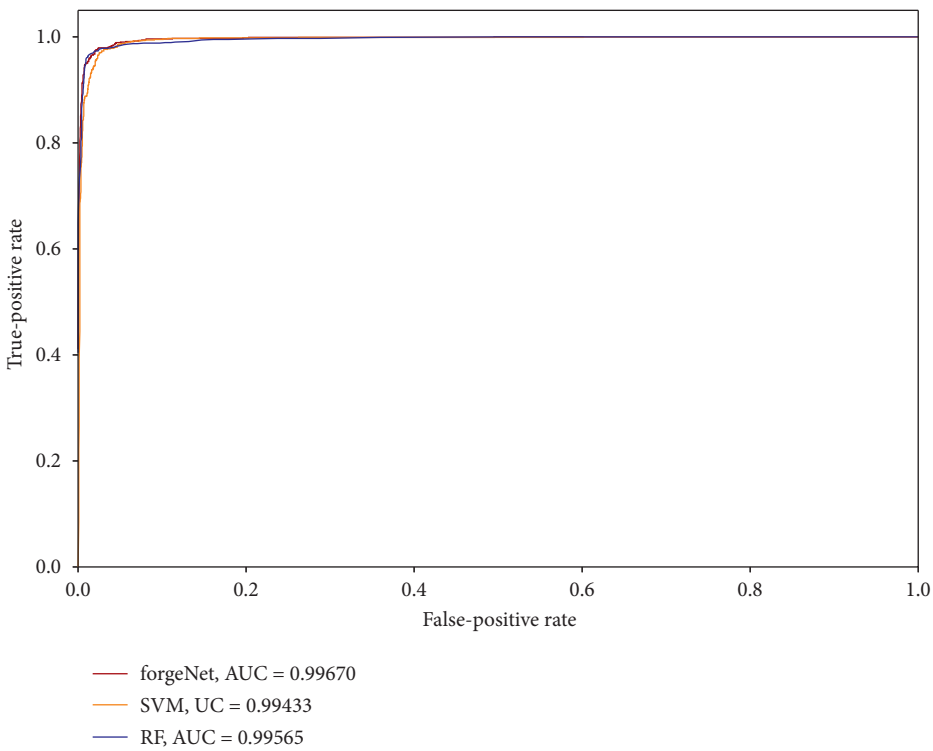

Figure 4: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 3 and 10-fold cross validation.

that, in terms of AUC, forgeNet is 0.34% higher than SVM and 0.11% lower than RF. Figure 10 proves that, in terms of AUC, forgeNet is 0.064% higher than SVM and little higher than RF. From the results of 3-fold cross validation, 5-fold cross validation and 8-fold cross validation, forgeNet has better ROC curves and higher AUC values than RF and SVM, which reveal that forgeNet could identify somatic mutations more accurately.

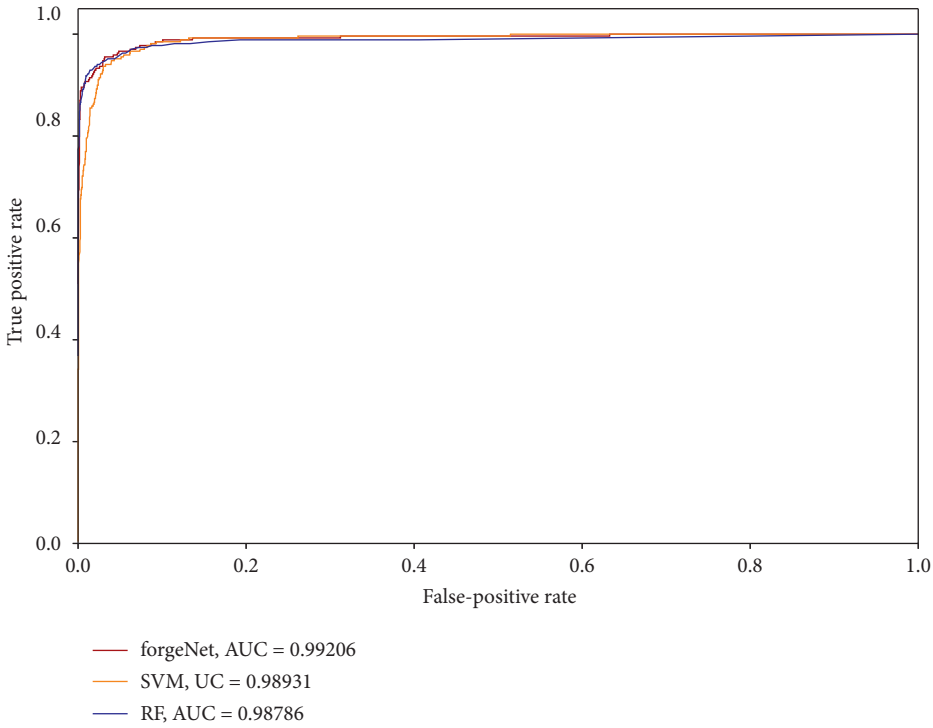

Figure 5: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 2 and 3-fold cross validation.

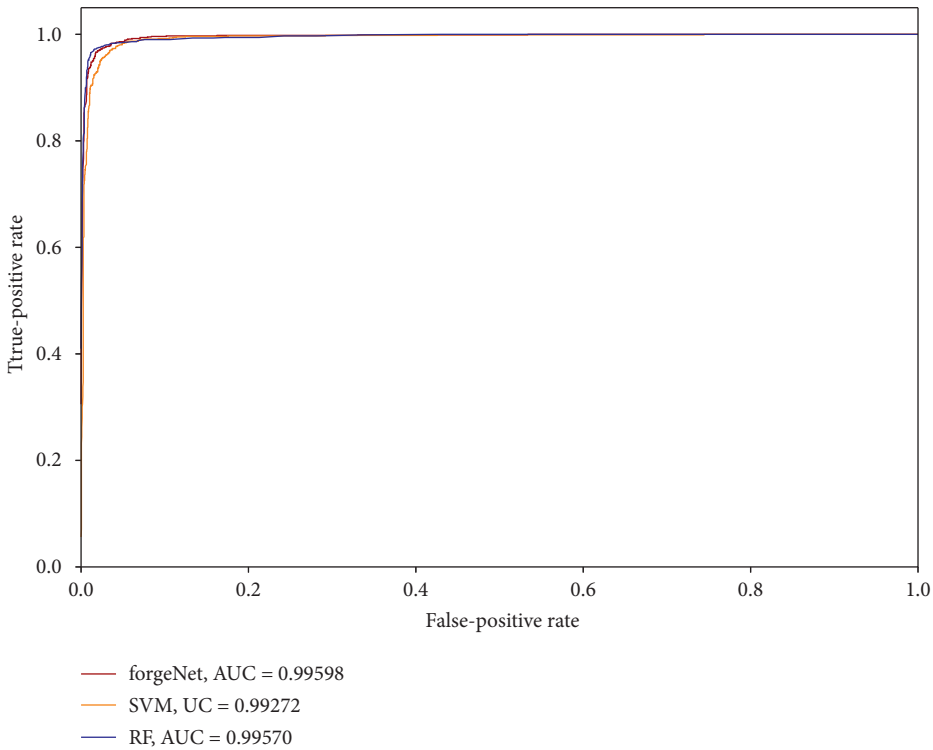

Figure 6: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 3 and 3-fold cross validation.

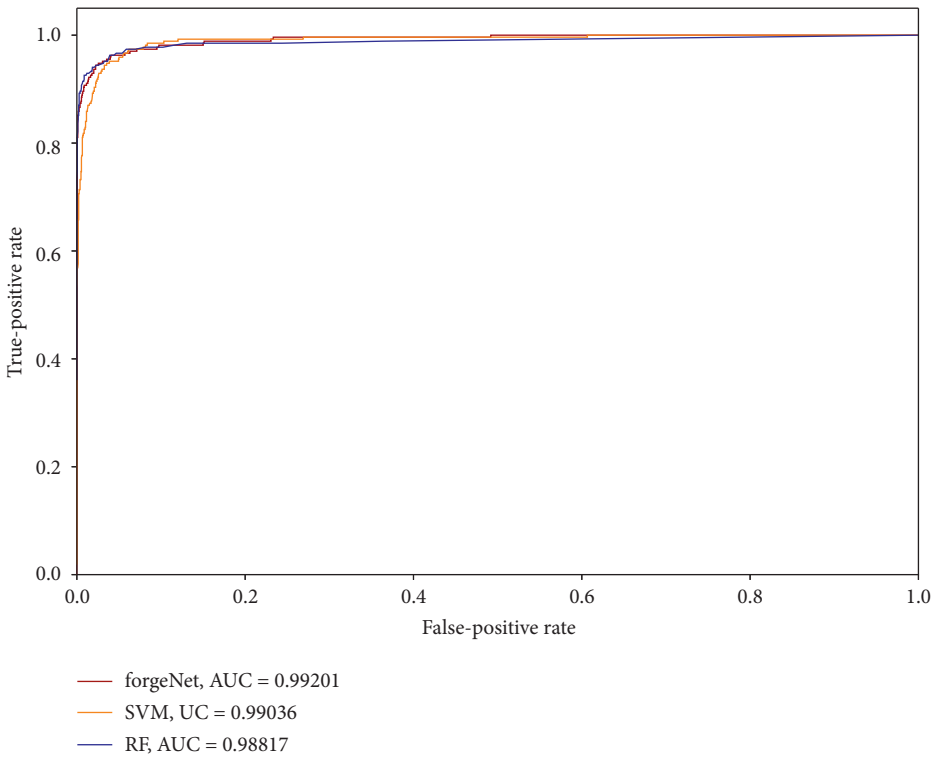

Figure 7: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 2 and 5-fold cross validation.

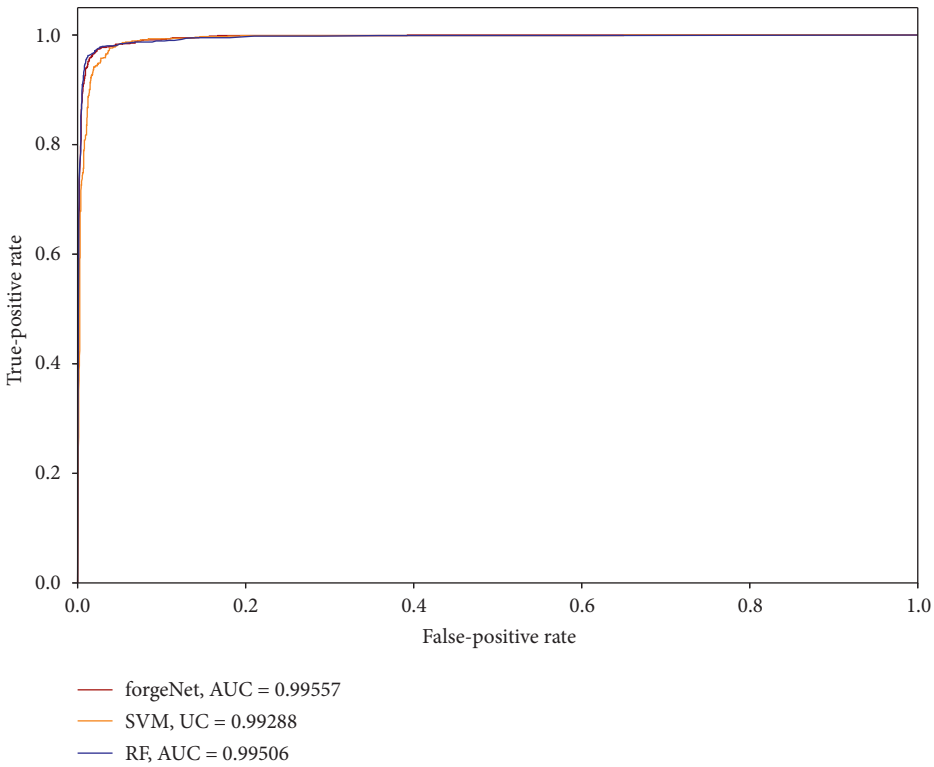

Figure 8: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 3 and 5-fold cross validation.

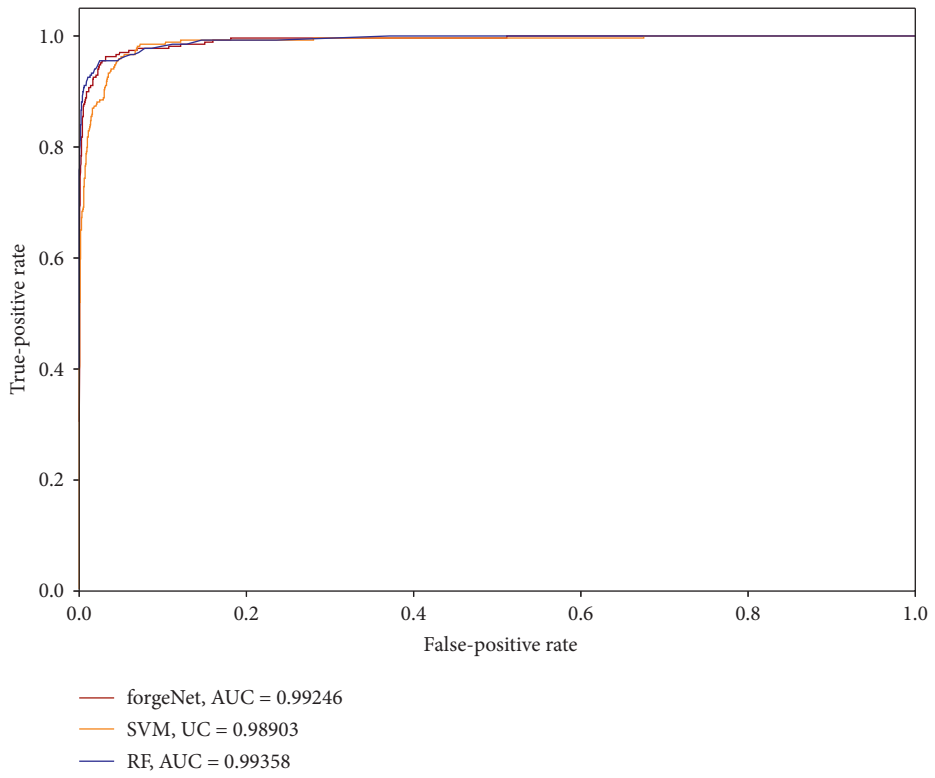

Figure 9: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 2 and 8-fold cross validation.

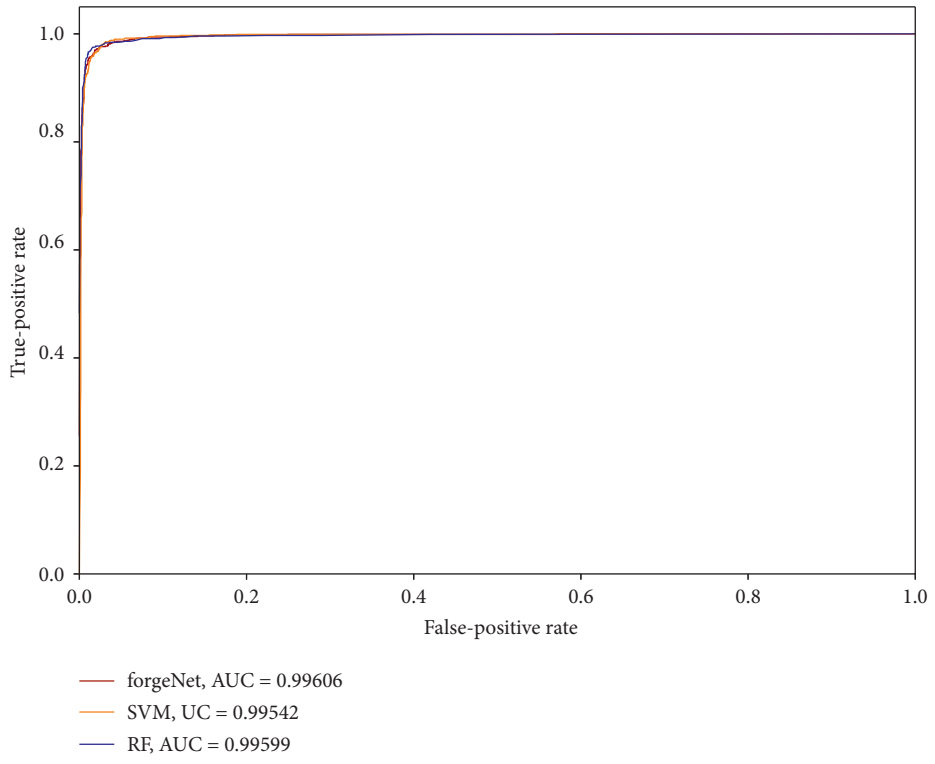

Figure 10: ROC curves and AUC values of somatic mutation detection by SVM, RF, and forgeNet with dataset 3 and 8-fold cross validation.

4. Conclusions

In this paper, a novel classifier, namely, forgeNet is utilized to improve the accuracy of somatic mutation identification. forgeNet contains two parts: the feature extractor part and neural network part, which are utilized to extract features and implement classification, respectively. Three real somatic mutation datasets are utilized to test the somatic mutation detection performances of forgeNet. Three-fold cross validation, 5-fold cross validation, 8-fold cross validation, and 10-fold cross validation are utilized. In terms of *SN*, *SP*, *Acc*, *MCC*, and $F1$, forgNet could identify more true somatic mutations, while random forest could identify more true nonsomatic mutations. The classification results reveal that forgeNet could make 0.05%–0.424% AUC improvement compared with support vector machines and random forest.

In the future, we will analyze the biological significance of somatic mutations in the process of classification. Also, the somatic mutations of different cancers will be classified and analyzed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

F. W. conceived the method. H. Q designed the method and wrote the main manuscript text. C. W. conducted the experiments. All authors reviewed the manuscript.

Acknowledgments

The authors acknowledge the funding received from the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001).

References

[1] E. Y. Chan, "Advances in sequencing technology," *Mutation Research*, vol. 573, no. 1-2, pp. 13–40, 2005.

[2] M. Pop and S. L. Salzberg, "Bioinformatics challenges of new sequencing technology," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[3] R. Elaine, *Mardis. The Impact of Next-Generation Sequencing Technology on Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[4] H. P. J. Buermans and J. T. Den Dunnen, "Next generation sequencing technology: advances and applications," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, 2014.

[5] X. Zhou, L. Ren, Y. Li, M. Zhang, Y. Yu, and J. Yu, "The next-generation sequencing technology: a technology review and future perspective," *Science China Life Sciences*, vol. 53, no. 1, pp. 44–57, 2010.

[6] N. Kumar, A. K. Mukhopadhyay, R. Patra et al., "Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India," *Journal of Bacteriology*, vol. 194, no. 21, pp. 5963-5964, 2012.

[7] D. Berner, M. Roesti, S. Bilobram et al., "De novo sequencing, assembly, and annotation of four threespine stickleback genomes based on microfluidic partitioned DNA libraries," *Genes*, vol. 10, no. 6, p. 426, 2019.

[8] J. H. Choo, C. P. Hong, J. Y. Lim et al., "Whole-genome de novo sequencing, combined with RNA-Seq analysis, reveals unique genome and physiological features of the amylolytic yeast Saccharomycopsis fibuligera and its interspecies hybrid," *Biotechnology for Biofuels*, vol. 9, no. 1, pp. 1–22, 2016.

[9] I. V. Bi, M. D. McMullen, H. Sanchez-Villeda et al., "Single nucleotide polymorphisms and insertion–deletions for genetic markers and anchoring the maize fingerprint contig physical map," *Crop Ence*, vol. 46, no. 1, pp. 12–21, 2006.

[10] R. T. Koehler, Z. Zhang, and A. R. Tobler, "Design of multiplexed oligionucleotide ligation assays for high throughput insertion-deletion polymorphism genotyping," *Cancer Research*, vol. 66, p. 696, 2006.

[11] D. R. Bentley, "Whole-genome re-sequencing," *Current Opinion in Genetics & Development*, vol. 16, no. 6, pp. 545–552, 2006.

[12] M. Martínez-Zapater José, R. Virginia, I. Ana et al., "High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology," *Bmc Genomics*, vol. 8, no. 1, p. 424, 2007.

[13] H. Scheffer, P. V. D. Vlies, M. Burton et al., "Two novel germline mutations of the retinoblastoma gene (RB1) that show incomplete penetrance, one splice site and one missense," *Journal of Medical Genetics*, vol. 37, no. 7, pp. 1–4, 2000.

[14] K. Masumura, N. Toyoda-Hokaiwado, A. Ukai et al., "Estimation of the frequency of inherited germline mutations by whole exome sequencing in ethyl nitrosourea-treated and untreated gpt delta mice," *Genes & Environment*, vol. 38, no. 1, p. 10, 2016.

[15] J. G. Ronquillo, C. Weng, and W. T. Lester, "Assessing the readiness of precision medicine interoperabilty: an exploratory study of the National Institutes of Health genetic testing registry," *Other*, vol. 24, no. 4, p. 918, 2017.

[16] M. Hollstein, D. Sidransky, B. Vogelstein, and C. Harris, "p53 mutations in human cancers," *Science*, vol. 253, no. 5015, pp. 49–53, 1991.

[17] S. Jones, X. Zhang, D. W. Parsons et al., "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–1806, 2001.

[18] K. Collett, "A basal epithelial phenotype is more frequent in interval breast cancers compared with screen detected tumors," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 5, pp. 1108–1112, 2005.

[19] N. Murphy, J. Mazda, and J. Gunter Marc, "Adiposity and gastrointestinal cancers: epidemiology, mechanisms and future directions," *Nature Reviews Gastroenterology & Hepatology*, vol. 15, pp. 659–670, 2018.

[20] B. Meier, N. V. Volkova, Y. Hong et al., "Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers," *Genome Research*, vol. 28, pp. 1371–1384, 2018.

[21] T. Mori, K. Miura, T. Aoki, T Nishihira, S Mori, and Y Nakamura, "Frequent somatic mutation of the MTS1/CDK4I (multiple tumor suppressor/cyclin-dependent kinase 4 inhibitor) gene in esophageal squamous cell carcinoma," *Cancer Research*, vol. 54, no. 13, pp. 3396-3397, 1994.

[22] A. Shinichiro, N. J. Sarlis, E. H. Oldfield et al., "Somatic mutation of TRbeta can cause a defect in negative regulation of TSH in a TSH-secreting pituitary tumor," *Journal of Clinical Endocrinology & Metabolism*, vol. 86, no. 11, pp. 5572–5576, 2013.

[23] N. Singh, D. K. Sahu, M. Goel, R. Kant, and D. K. Gupta, "Retrospective analysis of FFPE based Wilms' Tumor samples through copy number and somatic mutation related Molecular Inversion Probe Based Array," *Gene*, vol. 565, no. 2, pp. 295–308, 2015.

[24] N. Beije, J. C. Helmijr, M. J. A. Weerts et al., "Somatic mutation detection using various targeted detection assays in paired samples of circulating tumor DNA, primary tumor and metastases from patients undergoing resection of colorectal liver metastases," *Molecular Oncology*, vol. 10, no. 10, pp. 1575–1584, 2016.

[25] W. Bao, D. Wang, and Y. Chen, "Classification of protein structure classes on flexible neutral tree," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 5, pp. 1122–1133, 2017.

[26] B. Yang and W. Bao, "RNDEtree: regulatory network with differential equation based on flexible neural tree with novel criterion function," *IEEE Access*, vol. 7, pp. 58255–58263, 2019.

[27] J. Kubilius, S. Bracci, and H. O. D. Beeck, "Deep neural networks as a computational model for human shape sensitivity," *Plos Computational Biology*, vol. 12, no. 4, Article ID e1004896, 2016.

[28] W. Bao, C.-A. Yuan, Y. Zhang et al., "Mutli-features prediction of protein translational modification sites," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1453–1460, 2018.

[29] W. Bao, B. Yang, D. Li, Z. Li, Y. Zhou, and R. Bao, "CMSENN: computational modification sites with ensemble neural network," *Chemometrics and Intelligent Laboratory Systems*, vol. 185, pp. 65–72, 2019.

[30] J. Ding, A. Bashashati, A. Roth et al., "Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data," *Bioinformatics*, vol. 28, no. 2, pp. 167–175, 2011.

[31] S. Yuichi, S. Yusuke, C. Kenichi et al., "An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data," *Nuclc Acids Research*, vol. 41, no. 7, p. e89, 2013.

[32] D. C. Koboldt, D. E. Larson, and R. K. Wilson, "Using VarScan 2 for germline variant calling and somatic mutation detection," *Current Protocols in Bioinformatics*, vol. 44, no. 1, p. 44, 2013.

[33] S. M. E. Sahraeian, R. Liu, B. Lau et al., "Deep convolutional neural networks for accurate somatic mutation detection," *Nature Communications*, vol. 10, p. 1041, 2019.

[34] B. Yang and Y. Chen, "Somatic mutation detection using ensemble of flexible neural tree model," *Neurocomputing*, vol. 179, pp. 161–168, 2016.

[35] F. Dorri, S. Jewell, A. Bouchard-Cté et al., "Somatic mutation detection and classification through probabilistic integration of clonal population information," *Communications Biology*, vol. 2, 2019.

[36] Y. Kong and T. Yu, "forgeNet: a graph deep neural network model using tree-based ensemble classifiers for feature graph construction," *Bioinformatics*, vol. 36, no. 11, pp. 3507–3515, 2020.

[37] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, 2018.

[38] A. Sylvain, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[39] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Advances in Neural Information Processing Systems*, vol. 7, no. 10, pp. 231–238, 1995.

[40] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[41] I. Guyon, J. Weston, S. Barnhill et al., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[42] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2-3, pp. 18–22, 2002.

[43] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.