GenomeForest: An Ensemble Machine Learning Classifier for Endometriosis

Sadia Akter, PhD\textsuperscript{1}, Dong Xu, PhD\textsuperscript{1,2,3}, Susan C. Nagel, PhD\textsuperscript{4}, John J. Bromfield, PhD\textsuperscript{4}, Katherine E. Pelch, PhD\textsuperscript{4}, Gilbert B. Wilshire, MD\textsuperscript{5} and Trupti Joshi, PhD\textsuperscript{1,3,6}

\textsuperscript{1}Informatics Institute; \textsuperscript{2}Electrical Engineering and Computer Science; \textsuperscript{3}Christopher S. Bond Life Sciences Center, \textsuperscript{4}OB/GYN and Women's Health, University of Missouri, Columbia, MO; \textsuperscript{5}Boone Hospital Center, Columbia, MO; \textsuperscript{6}Health Management and Informatics, University of Missouri, Columbia, MO

Abstract

Endometriosis is a complex and high impact disease affecting 176 million women worldwide with diagnostic latency between 4 to 11 years due to lack of a definitive clinical symptom or a minimally invasive diagnostic method. In this study, we developed a new ensemble machine learning classifier based on chromosomal partitioning, named GenomeForest and applied it in classifying the endometriosis vs. the control patients using 38 RNA-seq and 80 enrichment-based DNA-methylation (MBD-seq) datasets, and computed performance assessment with six different experiments. The ensemble machine learning models provided an avenue for identifying several candidate biomarker genes with a very high $F_1$ score; a near perfect $F_1$ score (0.968) for the transcriptomics dataset and a very high $F_1$ score (0.918) for the methylomics dataset. We hope in the future a less invasive biopsy can be used to diagnose endometriosis using the findings from such ensemble machine learning classifiers, as demonstrated in this study.

Introduction

Endometriosis is a complex yet common gynecological disorder of reproductive-aged women. It is characterized by the presence of endometrial tissue outside of the uterine cavity. Endometriosis is a high impact disease, commonly associated with chronic pelvic pain and infertility. Therefore, it significantly impairs mental and physical quality of patient’s life and their work performance is seriously compromised. About 176 million women worldwide are suffering from endometriosis and about 8.5 million women solely in the North America\textsuperscript{1}. Endometriosis affects reproductive aged women (5-10%), women with subfertility (20-30%), and women with chronic pelvic pain and infertility (40-60%)\textsuperscript{2}. About 70% of teens who are suffering from pelvic pain are later diagnosed with endometriosis\textsuperscript{3}. In the U.S., endometriosis is a leading cause of hysterectomies (approximately 600,000 cases) performed every year\textsuperscript{4}. The total cost (direct and indirect) of endometriosis has been estimated at €30 billion in Europe and $22 billion in the U.S. each year, and direct costs have increased gradually\textsuperscript{5}.

Though laparoscopy is currently the gold standard diagnostic approach for endometriosis\textsuperscript{6}, it is an invasive procedure and may not be appropriate for all women with a history and physical examination indicative of endometriosis. There are many studies that assessed the diagnostic value of biomarkers for endometriosis in endometrial tissue, menstrual or uterine fluids and immunologic markers in blood or urine for clinical use as a diagnostic test for endometriosis; however, no reliable biomarkers were recommended\textsuperscript{7}. Due to the lack of reliable recommended biomarkers, the current diagnostic latency is on average 4 to 11 years\textsuperscript{8}. To reduce the sufferings and expenses related to the disease, an early intervention is essential. Studies have shown that endometriosis patients have an altered methylene and transcriptome, which could lead to the identification of biomarkers for developing a minimally invasive diagnostic technique for endometriosis\textsuperscript{9}.

Applications of machine learning methods on microarray expression data or next generation sequencing data have been advanced over the last several decades for discovery of biological patterns\textsuperscript{10}. For microarray expression data, use of both supervised and unsupervised machine learning methods have shown great success\textsuperscript{11}, including the application of: (a) clustering techniques such as hierarchical clustering and K-means clustering for identifying the groups of genes that share similar functions or expressions\textsuperscript{12}, and (b) disease vs. healthy classification tasks using various methods such as Decision Trees, Random Forests, Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Bayesian Networks\textsuperscript{13}. Although the application of machine learning classifiers on transcriptomics or methylomics data had limited success\textsuperscript{14,15}, the classification difference of gene expressions in transcriptomics data or the difference of DNA-methylation in methylomics data between disease vs. healthy can provide avenues for the development of endometriosis diagnostic method\textsuperscript{9,16}.

In our previous works, we have successfully demonstrated the application of various machine learning techniques for classifying the endometriosis patients vs. the control patients using both transcriptomics and methylomics data\textsuperscript{17,18}. In this work, we describe our new ensemble technique called GenomeForest based on chromosomal partitioning. We
systematically examined how well the newly developed ensemble technique perform in classifying endometriosis and control samples using both transcriptomics and methylomics data. The assessment was done from three different perspectives: (a) evaluation of classification performances of the GenomeForest ensemble classifier, (b) implication of three different normalization techniques, and (c) implication of differential analysis. The results were compared with the results from our prior work on the application of various machine learning techniques on the same dataset published elsewhere\textsuperscript{17,18}.

Methods

Subjects and Tissue Collection

The women participated in this study were aged between 18-49 years and all undergoing a laparoscopy procedure, either diagnostic laparoscopy for pain or infertility or seeking laparoscopic sterilization. Prior to surgery, the physician obtained informed consent following the IRB protocol. Endometrial biopsies were collected under general anesthesia prior to surgery. During laparoscopy, the physician thoroughly examined the peritoneal cavity and visually confirmed the presence or absence of endometriosis. Endometriosis patients had visually and histologically confirmed endometriosis. The control population were visually confirmed to be free of endometriosis. The tissue samples were processed for generating high-throughput mRNA transcriptomics data (RNA-Seq) and DNA methylomics data (MBD-seq). Our transcriptomics dataset includes 38 single-end RNA-seq samples (22 controls and 16 endometriosis) and the methylomics dataset includes 80 MBD-seq DNA methylation samples (36 controls and 44 endometriosis). More details can be found in our earlier publication\textsuperscript{18} on the same datasets.

Transcriptomics and Methylomics Data Preprocessing

We preprocessed the transcriptomics and methylomics data using several widely accepted bioinformatics tools. For transcriptomics data, we used FastQC, Cutadapt, Bowtie2, TopHat and HTSeq in different steps of the preprocessing. We used hg38 as reference genome. After getting the read count data from HTSeq, the rest of the analysis was performed using R packages. Low count genes were removed using the filtering criterion: keep the genes that have at least 1 count per million (cpm) reads mapped in at least \( n \) samples, where \( n \) is the smallest group size. For DNA-methylation data, we used FastQC, Cutadapt, Bowtie2, Samtools, Picard and R packages for preprocessing the data. We segmented the genome sequence into tiling windows of size 1,000 bases, which is widely used and recorded the number of reads that are mapped to each tiling windows/regions. Read count are the number of aligned reads that uniquely map to the hg38 reference genome. Very low count regions were filtered out using the filtering criterion: keep the regions that have non-zero counts per million (cpm) reads mapped in at least \( n \) samples, where \( n \) is the smallest group size. More details about the data preprocessing are available in our earlier publication\textsuperscript{18}.

The read count data was normalized using three different techniques: (a) logarithm of counts per million (logCPM) of trimmed mean of M values (TMM)\textsuperscript{19}, (b) Quantile normalization (qNorm)\textsuperscript{20}, and (c) Voom normalization (vNorm)\textsuperscript{19}. For differential analysis, a generalized linear model (GLM) followed by likelihood ratio test was applied using the edgeR package to identify differentially expressed genes (DEGs) in the transcriptomics data and differentially methylated regions (DMRs) in the methylomics data. The significance of the genes/methylated regions were defined by using an adjusted p-value cutoff set at 5% using the false discovery rate (FDR) method for multiple testing\textsuperscript{21}.

In the methylomics data analysis, our goal is to identify the methylated regions of interest (MROI) and find the nearby genes. Mapping of an MROI to the reference annotation file helped us to extract the nearest genes from that MROI. Our goal is to identify the genomic features such as the protein coding genes, long intervening noncoding RNA (lincRNA) genes, microRNA (miRNA) genes, Ribosomal ribonucleic acid (rRNA) genes, small nucleolar RNA (snoRNA) genes, and small nuclear RNA (snRNA) genes. The distance threshold for the MROI position to the genomic region was set to 10,000 bps. An MROI can be in the upstream/downstream region, or it can fall into a gene.

GenomeForest

In machine learning, an ensemble is a set of \( k \) base classifier models \((M_1, M_2, M_3, \ldots, M_k)\) for the purpose of creating an improved composite classification model \((M)\). A set of \( k \) training datasets \((D_1, D_2, D_3, \ldots, D_k)\) are created from the master dataset \((D)\), where \( M \) is created by training a classifier model on \( D_i \) \((1 \leq i \leq k)\). For classifying a new data tuple, the ensemble model \( M \) generates a class prediction based on the votes of the base classifiers. We developed an ensemble method called GenomeForest (Figure 1), in which each of the classifier is a decision tree classifier representing a classification model for each pair of chromosomes (up to 23) so that the collection is a forest representing the whole genome. C4.5\textsuperscript{22} is a popular algorithm for decision tree construction that uses entropy minimization or information gain for attribute selection criteria. We used an improved version of C4.5 (called C5.0/see5\textsuperscript{23}) for constructing the decision tree in this study. Confidence factor is used as a parameter for tree pruning.
in C5.0. The default value for confidence factor is 25% or 0.25. If the value of confidence factor is smaller than 0.25, it causes more pruning and vice versa.

In GenomeForest, given a whole genome sequencing dataset \( D \), up to 23 training datasets \( D_{Chr1}, D_{Chr2}, \ldots, D_{Chr22}, D_{ChrX} \) are created by partitioning the dataset, \( D \), in which a training dataset \( D_{ChrN} \) (where \( N \) in \( ChrN \) is the chromosome number 1, 2, 3, \ldots, 22, and \( X \)) contains the attributes corresponding to chromosome \( N \). A decision tree model \( DT_{ChrN} \) is trained by using a training dataset \( D_{ChrN} \). A composite prediction score \( PS \) for each class is calculated by using voting (score for each \( DT_{ChrN} \) model is 1) or weighted by various performance measures such as accuracy, sensitivity, specificity, precision, \( F_1 \) score, area under the curve (AUC) of receiver operating characteristics graph, and Matthews correlation coefficient (MCC). The final predicted class \( PC \) for a new data tuple is the class with the highest total prediction score. There are two class labels (endometriosis vs. control) in our transcriptomics and methylomics datasets. As such, the formula for calculating the composite or total prediction scores for the endometriosis and the control classes are presented in Formulas 1 and 2. Formula 3 identifies the final predicted class.

\[
\begin{align*}
\text{Total } PS_{endo} &= \sum_{N=1 \text{ to } 22 \text{ and } X} (\text{if } (PC(DT_{ChrN}) == endo) \text{ then } PS(DT_{ChrN}) \text{ else } 0) \quad (1) \\
\text{Total } PS_{cont} &= \sum_{N=1 \text{ to } 22 \text{ and } X} (\text{if } (PC(DT_{ChrN}) == cont) \text{ then } PS(DT_{ChrN}) \text{ else } 0) \quad (2) \\
\text{if } (\text{Total } PS_{endo} \geq \text{Total } PS_{cont}), \text{then } PC = endo \text{ else } PC = cont \quad (3)
\end{align*}
\]

Cross Validation and Model Performance

For model validation and comparing results between the experiments (described below), we applied the leave-one-out cross validation for computing the performance measures. This ensures two things: (1) the record used for model validation is not used for model construction, and (2) all records are used for model validation. This technique is useful for dataset with smaller number of samples such as in our study. The final model is constructed using all records. We computed several model performance measures: accuracy, sensitivity, specificity, precision, \( F_1 \) score, Matthews correlation coefficient (MCC), and area under the receiver operating characteristics curve (AUC); the leave-one-out cross validation approach was used for calculating these measures.

**Figure 1.** GenomeForest – an Ensemble of Decision Trees Representing Classifier for Each Chromosome for a Whole Genome Sequencing Dataset

**Machine Learning Experimental Approach**

We performed six different experiments using the GenomeForest ensemble classifier as shown in Table 1. Performance measures of each model were computed using the cross-validation approach described above. We used the default value of confidence factor (0.25) so that the decision tree is optimally pruned. For each of the GenomeForest experiment, we applied seven different criteria (such as accuracy, sensitivity, specificity, precision, \( F_1 \) score, MCC, and AUC) for ranking the decision tree models \( DT_{ChrN} \), used the highest ranked models \( (topN = 1, 2, 3, \ldots, 23) \) in the ensemble process, and eight different criteria for scoring (including accuracy, sensitivity, specificity, precision, \( F_1 \) score, MCC, AUC and voting). This experimental approach produced up to 1,288 \( (7 \times 23 \times 8) \) GenomeForest ensemble models for each category of GenomeForest experimental approach as listed in Table 1.
In experiment (1-3) (Table 1), we applied different normalization techniques on the raw read counts of genes/methylated regions and then applied GenomeForest. In experiment (4-6), differential analysis using GLM was performed first on each partitions of the dataset ($D_{chrN}$) to reduce features, such as genes in the transcriptomics data and genomic regions in the methylated data. After that, we applied different normalization techniques on the raw read counts of differential genes/methylated regions and then applied GenomeForest.

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>Experiment Name</th>
<th>Experiment Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) TMM + GenomeForest</td>
<td>(3) vNorm + GenomeForest</td>
<td>(5) qNorm + GLM + GenomeForest</td>
</tr>
<tr>
<td>(2) qNorm + GenomeForest</td>
<td>(4) TMM + GLM + GenomeForest</td>
<td>(6) vNorm + GLM + GenomeForest</td>
</tr>
</tbody>
</table>

The datasets were filtered for low read count genes for the transcriptomics datasets and for low read count methylated regions for the methylomics dataset. For the transcriptomics dataset, GenomeForest experiments were conducted in two scenarios: (a) all genes including protein coding, lincRNA gene, miRNA gene, rRNA gene, etc. are present in the dataset, and (b) only protein coding genes are present in the dataset. For the methylomics dataset, all methylated regions except lower read counts were present.

The results of the six experiments using the GenomeForest models were compared with the same set of experiments using the regular decision tree models as well as three machine learning models using an enhanced algorithm for detecting biomarkers named Biosigner\(^2\); the models in Biosigner include Partial Least Squares Discriminant Analysis (PLSDA), Random Forest (RF), and Support Vector Machine (SVM). Biosigner is an enhanced algorithm for detecting biomarkers. The details of the work on same dataset using decision tree and Biosigner experiments were published in Akter et al. (2019)\(^1\).

**Results**

**Transcriptomics GenomeForest Results**

After pre-processing of the 38 RNA-seq data, we created a dataset containing the read counts of 58,050 genes in which 18,852 genes were protein-coding. After filtering for low count genes, 14,154 genes were included in the dataset in which 11,687 of them were protein coding genes.

We applied the six experimental approaches using the ensemble-based GenomeForest algorithm on both protein coding and non-protein coding genes (denoted as “all genes” in this article) that includes 14,154 genes and on protein coding genes only that includes 11,687 genes. For each experimental approach, we created up to 1,288 GenomeForest models for various combinations of ranking matrices, value of topN and scoring matrices as described in the Method section. Out of the up to 1,288 GenomeForest models in each experimental approach, the performance measures of one of the best GenomeForest model are presented in Table 2. Within the six experimental approaches using all genes, the best performance was obtained for the “qNorm + GenomeForest” experiment by using the top 18 decision tree models ($DT_{chrN}$) ranked by the sensitivity measures and using precision as the scoring criteria. This experiment achieved the accuracy of 94.7%, sensitivity of 100%, specificity of 90.9%, precision of 88.9%, $F_1$ score of 0.941 and the MCC of 0.899. Within the six experimental approaches using protein coding genes, the best performance was obtained for the “vNorm + GenomeForest” experiment by using all 23 decision tree models ($DT_{chrN}$) and using $F_1$ score as the scoring criteria. This experiment achieved the accuracy of 97.4%, sensitivity of 93.8%, specificity of 100%, precision of 100%, $F_1$ score of 0.968 and MCC of 0.947. A total of 73 genes were identified by the individual decision tree models across all 23 pairs of chromosomes from the “vNorm + GenomeForest” experiment using the protein coding genes and three of these were differentially expressed and downregulated. We compared these 73 genes with the gene list found from the decision tree and Biosigner models in our previous study\(^1\); *NOTCH3*, *B4GALNT1* and *GTF3C5* were found common between GenomeForest and decision tree genes. All three genes were found downregulated in the differential analysis. Only *NOTCH3* was found common between GenomeForest and Biosigner.

<table>
<thead>
<tr>
<th>Gene Feature Set</th>
<th>Experiment Name</th>
<th>topN</th>
<th>Ranking Metric Name</th>
<th>Scoring Metric Name</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>$F_1$ Score</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>TMM + GenomeForest</td>
<td>23</td>
<td>NA</td>
<td>$F_1$ Score</td>
<td>0.895</td>
<td>0.938</td>
<td>0.864</td>
<td>0.833</td>
<td>0.882</td>
<td>0.792</td>
</tr>
<tr>
<td>All</td>
<td>qNorm + GenomeForest</td>
<td>18</td>
<td>Sensitivity</td>
<td>Precision</td>
<td>0.947</td>
<td>1.000</td>
<td>0.909</td>
<td>0.889</td>
<td>0.941</td>
<td>0.899</td>
</tr>
</tbody>
</table>

36
### Table 1: Performance Comparisons Using All Genes of the Transcriptomics Dataset

<table>
<thead>
<tr>
<th>Condition</th>
<th>Method</th>
<th>AUC Value</th>
<th>AUC Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>vNorm + GenomeForest</td>
<td>0.947</td>
<td>0.938</td>
<td>0.955</td>
</tr>
<tr>
<td>All</td>
<td>TMM + GLM + GenomeForest</td>
<td>0.737</td>
<td>0.813</td>
<td>0.682</td>
</tr>
<tr>
<td>All</td>
<td>qNorm + GLM + GenomeForest</td>
<td>0.816</td>
<td>0.813</td>
<td>0.818</td>
</tr>
<tr>
<td>All</td>
<td>vNorm + GLM + GenomeForest</td>
<td>0.816</td>
<td>0.813</td>
<td>0.765</td>
</tr>
<tr>
<td>Protein Coding</td>
<td>TMM + GenomeForest</td>
<td>0.947</td>
<td>1.000</td>
<td>0.909</td>
</tr>
<tr>
<td>Protein Coding</td>
<td>qNorm + GenomeForest</td>
<td>NA</td>
<td>1.000</td>
<td>0.909</td>
</tr>
<tr>
<td>Protein Coding</td>
<td>vNorm + GenomeForest</td>
<td>0.974</td>
<td>0.938</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### Figure 2. Performance Comparisons Using All Genes of the Transcriptomics Dataset

### Figure 3. Performance Comparisons Using the Protein Coding Genes of the Transcriptomics Dataset
Performance Comparisons of Models Using Transcriptomics Data

We have compared the performance of GenomeForest with our earlier machine learning classifier\textsuperscript{18} applied on the same dataset. A bar chart comparison of accuracy, sensitivity and specificity for experiments using all genes are presented in Figure 2. In this scenario, the “qNorm + GLM + GenomeForest” and “vNorm + GLM + GenomeForest” experiments have a balanced accuracy, sensitivity and specificity but does not outperform all of the experiments. The “qNorm + GenomeForest” experiment produced the highest accuracy and specificity among all the experiments and outperformed all of the experiments by F\textsubscript{1} score and MCC. A bar chart comparison of accuracy, sensitivity and specificity for experiments using the protein coding genes are presented in Figure 3. In this scenario, the “Biosigner (SVM)” method has a balanced accuracy, sensitivity and specificity but does not outperform all of the experiments. The “vNorm + GenomeForest” experiment produced the highest accuracy and sensitivity among all of the experiments and outperformed all experiments based on F\textsubscript{1} score and MCC. In both scenarios, GLM was useful for improving the overall performance in case of the decision tree application but GenomeForest was able to produce the best performance without using GLM.

Methylomics GenomeForest Results

We had 80 enrichment-based DNA-methylation (MBD-seq) samples where 77 samples met the quality control criteria (35 controls and 42 endometriosis). After pre-processing the data, we created a dataset containing the read counts of 3,088,281 methylated regions. After applying filtering criteria for lower read counts, 2,577,382 methylated regions were included in the dataset for further analysis.

We applied the six experimental approaches using the ensemble-based GenomeForest algorithm on the methylomics dataset that includes 2,577,382 methylated regions. For each experimental approach, we created up to 1,288 GenomeForest models for various combination of ranking matrices, values of topN and scoring matrices as described in the Method section. Out of the up to 1,288 GenomeForest models in each experimental approach, the performance measures of one the best GenomeForest model is presented in Table 3. Within the six experimental approaches, the best performance was obtained for both “qNorm + GenomeForest” and “vNorm + GenomeForest” experiments by using the top 22 decision tree models (DT\textsubscript{22}) ranked by the F\textsubscript{1} score and also using F\textsubscript{1} score as the scoring criteria. Both experiments achieved the accuracy of 90.9\%, sensitivity of 92.9\%, specificity of 88.6\%, precision of 90.7\%, F\textsubscript{1} score of 0.918 and the MCC of 0.817. A total of 109 MROIs were identified by the individual decision tree models across all 23 pairs of chromosomes from the “qNorm + GenomeForest” experiment. We found 24 genes within the distance of 10,000 bps from those 109 MROIs, in which the biotypes of the genes were as follows: protein coding (n=7), lincRNA (n=3), antisense (n=3), sense intronic (n=1), snoRNA (n=1), snRNA (n=2), and pseudogene (n=7). The location of the regions from the genes were upstream (n=15), downstream (n=4) or overlapping (n=5). We compared these 24 genes with the gene list found from the decision tree models\textsuperscript{18} and found MFS1DB to be common.

Performance Comparisons of Models Using Methylation Data

We have compared the performance of GenomeForest with our earlier machine learning classifier\textsuperscript{18} applied on the same dataset. A bar chart comparison of accuracy, sensitivity and specificity for experiments using the methylomics dataset are presented in Figure 4. The “qNorm + GLM + GenomeForest”, and “vNorm + GLM + GenomeForest” experiments have a balanced accuracy, sensitivity and specificity but does not outperform all of the experiments. Both the “qNorm + GenomeForest” and “vNorm + GenomeForest” experiments produced the highest accuracy, sensitivity and specificity among all the experiments and outperformed all of the experiments by F\textsubscript{1} score and MCC. In our earlier study, we have shown that GLM was useful to improve the overall performance in case of decision tree application\textsuperscript{18} but GenomeForest was able to produce the best performance without the help of GLM.

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>topN</th>
<th>Ranking Metric Name</th>
<th>Scoring Metric Name</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F\textsubscript{1} Score</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMM + GenomeForest</td>
<td>23</td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>0.870</td>
<td>0.929</td>
<td>0.800</td>
<td>0.848</td>
<td>0.886</td>
<td>0.740</td>
</tr>
<tr>
<td>qNorm + GenomeForest</td>
<td>22</td>
<td>F\textsubscript{1} Score</td>
<td>F\textsubscript{1} Score</td>
<td>0.909</td>
<td>0.929</td>
<td>0.886</td>
<td>0.907</td>
<td>0.918</td>
<td>0.817</td>
</tr>
<tr>
<td>vNorm + GenomeForest</td>
<td>22</td>
<td>F\textsubscript{1} Score</td>
<td>F\textsubscript{1} Score</td>
<td>0.909</td>
<td>0.929</td>
<td>0.886</td>
<td>0.907</td>
<td>0.918</td>
<td>0.817</td>
</tr>
<tr>
<td>TMM + GLM + GenomeForest</td>
<td>2</td>
<td>Specificity</td>
<td>Voting</td>
<td>0.831</td>
<td>0.811</td>
<td>0.771</td>
<td>0.822</td>
<td>0.851</td>
<td>0.659</td>
</tr>
<tr>
<td>qNorm + GLM + GenomeForest</td>
<td>7</td>
<td>AUC</td>
<td>AUC</td>
<td>0.818</td>
<td>0.810</td>
<td>0.829</td>
<td>0.850</td>
<td>0.829</td>
<td>0.636</td>
</tr>
<tr>
<td>vNorm + GLM + GenomeForest</td>
<td>1</td>
<td>F\textsubscript{1} Score</td>
<td>F\textsubscript{1} Score</td>
<td>0.792</td>
<td>0.810</td>
<td>0.771</td>
<td>0.810</td>
<td>0.810</td>
<td>0.581</td>
</tr>
</tbody>
</table>
Discussion
This work achieves our aim of broadly evaluating the newly developed ensemble approach, named GenomeForest, under different experimental scenarios. Also, we have successfully shown that it can improve model performances in classifying endometriosis and control samples using whole genome transcriptomics and methylomics data. Our newly proposed classifier outperforms various machine learning classifier algorithms (decision tree and Biosigner) on the same datasets that have been published earlier.

First, the ensemble technique can achieve a high classification accuracy. For whole genome sequencing data, we applied a logical partitioning approach based on each chromosome. We trained decision tree models on each chromosomal partition of the dataset and developed an ensemble approach for creating a composite collection of a forest model consisting of a set of decision trees representing the whole genome. We named this ensemble classification algorithm as GenomeForest. We experimented with various ranking and scoring techniques using GenomeForest and found that F1 score is best for both ranking and scoring. This model was able to outperform its counterpart.

Second, differential analysis using the GLM is widely used to identify the DEGs from the transcriptomics datasets and DMRs from the methylomics datasets. In our previous work, we have shown that differential analysis was useful for improving the performance of decision tree application by reducing the features (genes/genomic regions). However, in this work, we have shown that GenomeForest was able to produce the best performance without the help of differential analysis. We also evaluated different normalization techniques as a classifier’s performance may vary depending on the normalization techniques. We found that qNorm performed the best when all genes were considered and vNorm performed the best when only the protein-coding genes were considered in the transcriptomics dataset. For the methylomics dataset, both qNorm and vNorm normalizations performed the best in the GenomeForest application.

Third, the ensemble machine learning classifiers can be trained for creating highly accurate models for classifying endometriosis with high sensitivity and specificity thus creating the opportunity for precision medicine application for endometriosis. Mainly because of the complexity in diagnosis techniques, there is a delay from symptom onset to diagnosis ranging from 4 to 11 years which is very high. The ensemble machine learning models in this study achieved a very high F1 score; a near perfect F1 score (0.968) for the transcriptomics dataset and a very high F1 score (0.918) for the methylomics dataset. Although the predicted markers require further experimental and clinical validations, we hope in the future a less invasive biopsy can be used to diagnose endometriosis using ensemble machine learning classifiers-based findings as demonstrated in this study.

Fourth, we experimented if ensemble of a few top ranked chromosomes (ranked by different performance measures) could classify the disease samples from the controls. We observed that the best ensemble model used various numbers of decision tree models in different scenarios: (a) for the transcriptomic dataset using all genes, the best ensemble model was created using top 18 chromosomes ranked by sensitivity and scored by precision, (b) for the transcriptomic...
dataset using protein coding genes only, the best ensemble model was created using all chromosomes, (c) for the methylation data, the best model was created using the top 22 chromosomes ranked and scored by F1. Therefore, we concluded that instead of using few chromosomes, ensemble of most of the chromosomes in the GenomeForest could give us the best classification model.

Fifth, GenomeForest can assist in the identification of candidate biomarkers of endometriosis using transcriptomics and methylomics data. In our previous study, we discussed the candidate biomarker genes of endometriosis extracted from decision tree and Biosigner models. All the machine learning models (GenomeForest, decision tree and Biosigner) have identified NOTCH3 as a candidate biomarker. It is also differentially expressed and downregulated in our study. The NOTCH3 signaling may play a major role in onogenesis, tumor maintenance, and resistance to chemotherapy. NOTCH3 is associated with breast cancer development and pancreatic ductal adenocarcinoma (PDAC), lung carcinogenesis and endometrial carcinoma. Dysregulation and decrease in NOTCH signaling pathway is also associated with endometriosis. B4GALNT1 and GTF3C5 were identified by the decision tree and GenomeForest experiments. Trimarchi et al. (2017) identified B4GALNT1 to be related to endometrial cancer. B4GALNT1 is associated with two pathways: Glycosphingolipid biosynthesis - ganglio series and Sphingolipid metabolism, and diseases named Spastic Paraplegia 26 and autosomal recessive. GTF3C5 was reported as differentially expressed between endometrioid endometrial cancer and non-endometrioid endometrial cancer. Some other candidate biomarker genes are also related to different types of cancer. For example, ZBTB8A may be involved in gastric carcinoma, gastric adenocarcinoma cell differentiation, cancer invasion and metastasis. Aghajanova et al. (2011) identified H1FX to be differentially expressed in the comparison of severe versus mild endometriosis samples in the mid-secretory phase endometrium. Other known functions of H1FX are Cancer, Cell-To-Cell Signaling and Interaction, and Skeletal and Muscular Disorders. AIMP1 is a cytokine that is specifically induced by apoptosis, and involved in the control of angiogenesis, inflammation, and wound healing. It is also involved in the stimulation of inflammatory responses after proteolytic cleavage in tumor cells. Baek et al. (2018) identified KLC4 to be associated with human lung cancer cell lines. XRCC2 promotes colorectal cancer cell growth, regulates cell cycle progression, and apoptosis and mutations in XRCC2 can increase the risk of breast cancer. The aberrant expression of MED19 is involved in tumorigenesis and it promotes the proliferation of breast cancer. Tamaresis et al. (2014) identified NRXN3 to be differentially expressed in severe vs. mild endometriosis. CELF2 was reported as a putative tumor suppressor gene in colon cancer. IL17RA plays a pathogenic role in many inflammatory and autoimmune diseases. Bunch et al. (2011) reported that the expression of PGRMC1 is significantly decreased in the endometrium of women with endometriosis. Fang et al. (2011) suggested that abnormal HRH4 expression plays a role in the progression of colorectal cancers.

Lastly, the finding of many cancer and tumor associated genes using GenomeForest approach is consistent with our previous findings using other machine learning classifiers. Several studies investigated the relationship of endometriosis with cancer. Sato et al. (2000) and Thomas et al. (2000) have found some cancer associated mutations in endometriotic lesions and significant shared genetic correlation in both endometrial cancer and endometriosis. Some other studies have found that endometriosis patients are at a higher risk of developing several malignancies: ovarian cancer, breast cancer, renal cancer, thyroid cancer and brain tumor. Both cancer and endometriosis have some similar characteristics: metastasis, angiogenesis and resistance to apoptosis. However, only endometriosis is considered to be a benign condition. More genomic studies are needed to investigate the association of endometriosis with cancer.

In summary, this study demonstrated that GenomeForest, an ensemble machine learning classifier, is a robust and reliable approach for classifying endometriosis using transcriptomics or methylomics data. We concluded that an appropriate GenomeForest diagnostic pipeline for endometriosis should use (a) either transcriptomics or methylomics data, (b) vNorm for protein-coding genes, qNorm for all genes and either qNorm or vNorm for the methylomics data, (c) chromosomal partitioning with ensemble of decision trees for greatest increase in classification performance, (d) no differential analysis is necessary for feature reduction, and (e) F1 score for both ranking of the individual models and generating the composite score. The conclusion was made based on the use of endometriosis classification in this study. Further study is needed to generalize the results across multiple disease classification cases. Also, development of prediction classification models for endometriosis and other diseases using integrated multi-omics data would be an interesting investigation.

Data Availability
Acknowledgement
The biological samples were collected in collaboration with the University of Missouri and Boone Hospital, Columbia, MO, and University of California, San Francisco, CA. Some tissue samples were provided by the NIH SCCPIR Human Endometrial Tissue Bank and DNA Bank at UCSF (NIH grant HD055764-06). We are thankful to Stacy Syrcle, Sarah Crowder, Danny J. Schust, and Breton Barrier for help with sample collection. This work was supported by NIEHS R21 ES020039, University of Missouri Institute for Clinical and Translational Science, NIH/National Center for Advancing Translational Sciences (NCATS) grant UL1TR002345, and University of Missouri Research Council.

References