

Identification of Important Diagnostic Genes in the Uterine Using Bioinformatics and Machine Learning

Hossein Valizadeh Laktarashi¹, Milad Rahimi², Kimia Abrishamifar³, Ali Mahmoudabadi⁴, Elham Nazari^{5*} 

Received: 17 Aug 2024

Published: 6 Jan 2025

Abstract

Background: Uterine corpus endometrial cancer (UCEC) is known as the sixth most common cancer in the world. Advances in bioinformatics and deep learning have provided the 2 tools for screening large-scale genomic data and discovering potential biomarkers indicative of disease states. This study aimed to investigate the identification of important genes for diagnosis and prognosis in the uterus using bioinformatics and machine learning algorithms.

Methods: RNA expression profiles of UCEC patients were analyzed to identify differentially expressed genes (DEGs) using deep learning techniques. Prognostic biomarkers were assessed through survival curve analysis utilizing COMBIO-ROC. Additionally, molecular pathways, protein-protein interaction (PPI) networks, co-expression patterns of DEGs, and their associations with clinical data were thoroughly examined. Ultimately, diagnostic markers were determined through deep learning-based analyses.

Results: According to our findings, MEX3B, CTRP2 (C1QTNF2), and AASS are new biomarkers for UCEC. The evaluation metrics demonstrate the deep learning model's (DNN) efficacy, with a minimal mean squared error (MSE) of 5.1096067E-5 and a root mean squared error (RMSE) of 0.007, indicative of accurate predictions. The R-squared value of 0.99 underscores the model's ability to explain a substantial portion of the variance in the data. Thus, the model achieves a perfect area under the curve (AUC) of 1, signifying exceptional discrimination ability, and an accuracy rate of 97%.

Conclusion: The GDCA database and deep learning algorithms identified 3 significant genes —MEX3B, CTRP2 (C1QTNF2), and AASS—as potential diagnosis biomarkers of UCEC. Thus, identifying new UCEC biomarkers has promise for effective care, improved prognosis, and early diagnosis.

Keywords: Uterine Corpus Endometrial Carcinoma, Deep learning, Biomarker, Bioinformatic Analysis, UCEC

Conflicts of Interest: None declared

Funding: This research was funded and supported by Shahid Beheshti University of Medical Sciences.

***This work has been published under CC BY-NC-SA 4.0 license.**

Copyright© Iran University of Medical Sciences

Cite this article as: Valizadeh Laktarashi H, Rahimi M, Abrishamifar K, Mahmoudabadi A, Nazari E. Identification of Important Diagnostic Genes in the Uterine Using Bioinformatics and Machine Learning. *Med J Islam Repub Iran*. 2025 (6 Jan);39:4. <https://doi.org/10.47176/mjiri.39.4>

Introduction

Uterine corpus endometrial carcinoma (UCEC) is the fourth most common cancer among women in the United States and the sixth most prevalent cancer globally (1, 2). UCEC is one of the few cancers whose occurrence and

mortality rates are increasing in the United States, and to some extent, it reflects the prevalence of overweight and obesity since the 1980s (3). In 2020, the number of individ-

Corresponding author: Dr Elham Nazari, Elham.Nazari@sbmu.ac.ir

¹ Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

² Department of Health Information Technology, School of Allied Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran

³ School of Health Management and Information Science, Shiraz University of Medical Sciences, Shiraz, Iran

⁴ Department of Medical Genetics, Afzalipoor Faculty of Medicine, Kerman University of Medical Sciences, Kerman, Iran

⁵ Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

↑What is “already known” in this topic:

Bioinformatics analysis explored prognosis-associated biomarkers for estrogen-independent uterine corpus endometrial carcinoma. In summary, 4611 (messenger ribonucleic acid) mRNA, 3568 long non-coding RNAs (lncRNAs), and 47 microRNAs (miRNAs) were differentially expressed between endometrial cancerous and normal endometrial tissues.

→What this article adds:

Important diagnostic genes in the uterine were identified using bioinformatics and machine learning. According to findings, MEX3B, CTRP2 (C1QTNF2), and AASS are new biomarkers for uterine corpus endometrial cancer. The evaluation metrics demonstrate the deep learning model's efficacy, with a minimal mean squared error (MSE) of 5.1096067E-5 and a root mean squared error (RMSE) of 0.007, indicative of accurate predictions.

uals with UCEC around the world was 417,367. It is predicted that in 2023, almost 66,200 individuals in the United States will be diagnosed with UCEC (4). Further, it is predicted that the number of people suffering from this cancer will increase by 52.7% by 2040, and the death rate will reach 70.6% (1). According to histopathological features, uterine cancer is classified into 2 categories: endometrial cancer (very common) and uterine sarcoma (rare). UCEC includes about 80% of all cases with the disease, which has a survival rate of about 90% (1). Early detection of this cancer is exceptionally vital in managing its treatment. Improving decision-making and treatment management requires the identification of biomarkers related to this disease (5). In recent decades, methods such as decision fusion techniques, machine learning, and bioinformatics techniques have been used to analyze and process big data (6-8). Many studies have been introduced to identify biomarkers for the prognosis and diagnosis of UCEC. To identify these biomarkers in high-throughput data, bioinformatics and artificial intelligence approaches have been widely used (1, 9-11).

Zagidullin et al demonstrated how 3 visually interpretable models can contribute to generating new research hypotheses. By analyzing the decision path structure of the (Optimal Stopping Time) OST, they identified L1 cell adhesion molecule expression and estrogen receptor status as key risk factors in the p53 abnormal Endometrial Carcinoma (EC) subgroup (12).

Zhao et al utilized bioinformatics analysis to identify key genes and potential prognostic biomarkers for uterine cancer. Their study included gene set enrichment analysis (GSEA) and an evaluation of prognostic values and molecular mechanisms. They identified 28 upregulated and 94 downregulated genes across 4 gene expression omnibus (GEO) datasets after gene fusion. Gene ontology analysis indicated that these differentially expressed genes (DEGs) primarily regulate transcription and cell proliferation. Protein-protein interaction (PPI) analysis highlighted 10 hub genes—JUN, UBE2I, GATA2, WT1, PIAS1, FOXL2, RUNX1, EZR, TCF4, and NR2F2—with the highest dependency scores. Among these, the expression patterns of 9 genes, excluding UBE2I, aligned with mRNA levels from The Cancer Genome Atlas (TCGA) data. Furthermore, FOXL2, TCF4, and NR2F2 were significantly associated with uterine cancer prognosis, with their low expression correlating with poorer outcomes (13).

To identify early biomarkers of UCEC, researchers identified sets of possible candidates (ToppGene) using DisGenNET and gene expression databases and a prioritization algorithm. In the next step, PPI network analysis and survival analysis were used to analyze the data. They identified a total of 10 genes, among which the target protein Xklp2 (TPX2) was the most promising independent prognostic marker in the first stage of UCEC (1). Also, to predict the worst prognosis of type 2 UCEC, weighted gene co-expression network analysis (WGCNA), a combination of ceRNA regulatory network analysis, functional enrichment analysis, PPI network construction, and survival analysis identified 5 prognostic biomarkers—LINC02418, RASGRF1, GCNT1, LEF1, and NKD1 (11). In another

study, prognostic analysis showed that Cyclin E1 was significantly associated with worse overall survival in patients with UCEC. The hub genes and differentially expressed miRNAs identified in this study have demonstrated potential as prognostic biomarkers for UCEC and could serve as molecular targets for therapeutic interventions. These findings suggest promising avenues for utilizing molecular signatures to improve prognostication and treatment strategies in UCEC management (14).

The present study aimed to identify diagnostic biomarkers for UCEC. This research differentiates from other similar research in using a combination of DNN and bioinformatics analysis.

Methods

Data Collection

Gene data for UCEC patients, along with key clinical features such as sex, tumor stage, TNM (Tumor, Node, Metastasis) classification, and survival information, were retrieved from the GDAC dataset available at <https://gdac.broadinstitute.org/>. These well-curated datasets formed the basis for subsequent clinical analyses.

Preprocessing of Data and Identification of DEGs

Advanced machine learning techniques were used to discover novel genes, with careful awareness of the critical steps of normalization and filtering in the data analysis pipeline. Before using deep learning algorithms for the RNA data, thorough preprocessing was performed, which included filtering and normalization processes. Initially, duplicate genes and samples were systematically removed through custom filter scripts in R programming, ensuring that the dataset was free from redundancy and noise that could skew the analysis.

Normalization is a crucial step in the preprocessing pipeline, as it adjusts for variations between samples that are not related to the biological differences of interest. This step ensures that the expression levels of genes are comparable across all samples, which is essential for accurate downstream analysis. A total of 20,532 genes were standardized utilizing the widely recognized Limma and EdgeR packages within the R software framework. These packages are well-suited for handling RNA-seq data, as they offer robust statistical methods to correct for biases and scale differences, thereby enhancing the reliability of the subsequent analysis. After normalization, the DEGs were carefully identified by applying strict criteria. The logarithmic change threshold ($|FC|$) of ≥ 1.5 was chosen as a key parameter to ensure that only genes with a substantial change in expression levels were considered significant. This threshold strikes a balance between sensitivity and specificity, allowing for the detection of meaningful biological changes while minimizing the inclusion of genes with trivial or negligible expression differences. The significance level was set at $P < 0.05$ and was applied to ensure that the observed changes were statistically significant and not due to random variation. All analyses and visualization activities were implemented seamlessly by applying R software (Version 4.2.3), ensuring robustness and accuracy at each step of the process. This comprehensive approach to preprocessing,

normalization, and filtering enhances the validity of the findings and supports the discovery of novel genes with potential biological importance.

Identifying Predictive Markers

A bioinformatic analysis was conducted to identify DEGs as significant markers in UCEC using machine learning techniques. Specifically, we used DNN, a deep learning model renowned for its high performance in binary classification tasks, particularly in case-control prediction scenarios.

Deep learning

-Deep learning, a branch of artificial intelligence, has revolutionized numerous activities by enabling computers to learn from large many of data and make predictions or decisions without explicit programming. Deep learning mimics how the human brain processes information through artificial neural networks (15).

In deep learning, these neural networks consist of multiple layers of interconnected nodes called neurons. Each layer processes information and extracts features from the data, gradually learning more abstract representations as information passes through successive layers. This hierarchical representation allows deep learning models to automatically discover intricate patterns and relationships in complex datasets (16).

Deep learning represents a powerful subset of artificial intelligence, characterized by its ability to autonomously learn complex patterns and representations from vast amounts of data and an advanced DNN-based predictive model was designed to anticipate exhaustion behavior observed in textile dyeing processes. DNN, a machine learning technique built on an artificial neural network (ANN), emulates the structure and functions of the human neural network. An ANN comprises layers—including an input layer, hidden layers, and an output layer. Systems with three or more hidden layers are designated as DNNs. In tackling regression or classification issues, a linear estimation function like $y = wTx + b$ is employed. To address non-linear problems, a DNN integrates an activation function with the linear estimation function (17).

Implementation and Evaluation

The implementation of DNN was conducted using the Python programming language version 3.7. Several essential packages were utilized for this purpose—including Pandas, NumPy, TensorFlow, Keras, and PyTorch. These packages provide a robust framework for data manipulation, numerical computation, and building and training deep learning models.

The developed models underwent optimization using the training data and were subsequently independently evaluated using the test data. A train/test ratio of 70/30 was selected as the most effective for deep learning methods.

Performance metrics were employed to assess the effectiveness of the methods in identifying important genes. Five key indicators were considered for evaluation:

Accuracy

This metric evaluates the proportion of true positives and true negatives in the classification process, representing the degree of agreement between predicted and actual values.

F1 Score

The F1 Score is a metric used to assess the balance between precision and recall, particularly useful for evaluating classification models on imbalanced datasets.

Area Under the Receiver Operating Characteristic Curve

Area under the curve (AUC) is a metric that quantifies the potential of a classification model to differentiate between classes. The receiver operating characteristic (ROC) curve plots the true positive rate as opposed to the false positive rate, and the AUC appears for the area under this curve (18).

Confusion Matrix

The confusion matrix provides a tabular summary of the performance of a classification model, indicating the counts of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP).

R2 Score

The R2 Score (coefficient of determination) is commonly used to assess the goodness of fit of regression models. In this context, it is employed to evaluate the performance of models from a feature selection perspective.

These metrics collectively provide a comprehensive evaluation of the deep learning methods, offering insights into their accuracy, robustness, and discriminative ability.

Functional and Pathway Enrichment Analysis and PPI Network Analysis

Functional enrichment analysis and identification of pivotal pathways in the signature of DEGs were carefully performed using the Cluster Profiler package in R, with a strictly adjusted threshold of $P < 0.05$. In addition, to gain deeper insight into selected prognostic genes, comprehensive annotations and visualizations were performed using 2 widely known databases: Gene Ontology and Kyoto Encyclopedia of Genes and Genomes. The PPI network for the DEGs was visualized using the STRING biological database, available at <https://string-db.org/> (18). These interactions are critical for understanding cellular pathways and functional genomics. To ensure the accuracy and relevance of the identified interactions, a statistically significant interaction score threshold >0.4 was applied.

Correlation Between DEGs and Clinical Data

A completeness analysis was conducted to search the relationship between DEGs and clinical data—including age, malignant mass size, lymph node involvement, distant metastasis, and stage. Using the R programming language, specifically leveraging the ggcorrplot package and cor function, a correlation matrix and Spearman correlation

were employed to analyze 55 DEGs in conjunction with the aforementioned clinical data. This approach facilitated a thorough study into potential associations between gene expression patterns and clinical characteristics.

Identifying Diagnostic Markers

A generalized linear model was applied alongside combined ROC curve analysis to assess diagnostic performance and develop diagnostic models. Essential metrics—such as sensitivity, specificity, cutoff value, positive predictive value, negative predictive value, and the area under the ROC curve, were thoroughly evaluated to determine the discriminatory capability of individual or combined biomarkers. This analysis was conducted using the “COM-BIO-ROC” package within the R environment.

Validation of Biomarkers Gene Expression

The expression levels of the candidate genes in UCEC patients were validated using data from the Global Data Assembly Centers (GDAC), accessible at <https://gdac.broadinstitute.org/>. Further verification was performed using GEO datasets—including GSE119041, GSE17025, GSE115810, GSE36389, and GSE25405. Validation data for UCEC patients were obtained from this online resource, followed by preprocessing to ensure data quality and consistency.

Results

Patient Characteristics

This study presents a comprehensive analysis of a dataset comprising 548 individuals with 2533 genes. According to Table 1, descriptive statistics revealed a varied demographic profile, with patients aged between 0 and 90 years (mean, 63.58; SD, 12.075) at initial pathologic diagnosis. Notably, the dataset included patients with a broad range of follow-up periods (mean, 646.06; SD, 760.686), exhibiting a maximum follow-up duration of 5691 days. Moreover, patient height demonstrated considerable diversity, ranging from 0 to 183 cm (mean, 152.66; SD, 36.979). Categorical variables elucidate the distribution of clinical and demographic characteristics, with most patients exhibiting positive neoplasm cancer status (78.6%) and clinical stages 1 (62.4%), followed by stages 3 (22.6%) and 2 (9.5%). Additionally, the ethnic composition primarily comprises individuals identified as belonging to category 1 (68.8%). Further exploration of patient vital status underscores a predominantly surviving cohort (91.8%). The dataset also includes cases categorized as controls (99.3%). These findings offer valuable insights into the heterogeneity of patient characteristics and clinical factors, forming a foundational basis for subsequent analyses and investigations in oncological research.

Identification of DEGs

The dataset, consisting of 480 patients, was acquired from the GDAC database, and the creation of the data frame involved utilizing the cleaning and preprocessing methods.

Subsequently, 20531 DEGs were identified after normalization, and a heatmap was generated for visualization (Figure 1, A and B). After the normalization operation, a total

Table 1. Clinicopathological Characteristics of Patients

| Clinicopathological Variables | | No. of Patients (%) /mean ± SD |
|--|----|-----------------------------------|
| Patients | | 548 |
| 1. Mean age, years (mean ± SD) | 1. | 63.58 ± 12.075 |
| 2. Mean height, cm (mean ± SD) | 2. | 152.66 ± 36.97 |
| Race | | |
| 1. NA | 1. | 32 (5.8) |
| 2. white | 2. | 374 (68.2) |
| 3. black or African American | 3. | 109 (19.9) |
| 4. Asian | 4. | 20 (3.6) |
| 5. American Indian or Alaska Native | 5. | 4 (0.7) |
| 6. native Hawaiian or another pacific islander | 6. | 9 (1.6) |
| Patient ethnicity | | |
| 1. NA | 4. | 156 (28.5) |
| 2. Not Hispanic or Latino | 5. | 377 (68.8) |
| 3. Hispanic or Latino | 6. | 15 (2.7) |
| Stage | | |
| 1. Stage i | 1. | 342 (62.4) |
| 2. Stage ii | 2. | 52 (9.5) |
| 3. Stage iii | 3. | 124 (22.6) |
| 4. Stage iv | 4. | 30 (5.5) |
| Patient cancer status | | |
| 1. NA | 1. | 38 (6.9) |
| 2. No | 2. | 431 (78.6) |
| 3. Yes | 3. | 79 (14.4) |
| Vital status | | |
| 1. Dead | 1. | 503 |
| 2. Alive | 2. | 45 |

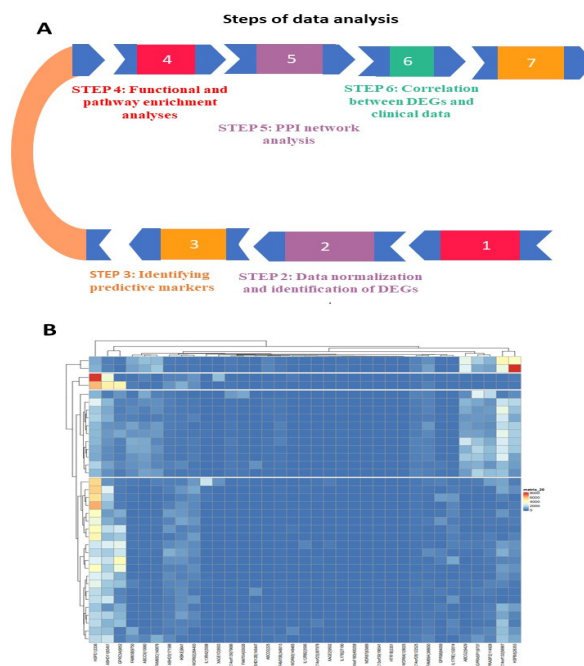


Figure 1. A The overall workflow, B Then heatmap of DEGs of UCEC was drawn by R software

of 1047 genes finally remained.

Identifying Predictive Genes

The deep learning model in this study, identified as DNN, exhibits highly promising performance in a binary classifi-

cation task focused on case-control prediction. The evaluation metrics demonstrate the model's efficacy, with a minimal mean squared error (MSE) of 5.10×10^{-10} and a root mean squared error (RMSE) of 0.007, indicative of accurate predictions. The R-squared value of 0.99 underscores the model's ability to explain a substantial portion of the variance in the data. Furthermore, the model achieves a perfect AUC of 1, signifying exceptional discrimination ability, and an accuracy rate of 97%. The precision-recall AUC also attains a maximum value of 1. The confusion matrix reveals impeccable classification performance, with zero errors in predicting both positive and negative cases. The gains/lift table demonstrates the model's efficiency in identifying positive cases across different thresholds. Variable importance highlights significant features influencing the model, contributing to its robust predictive capabilities. The status of neuron layers provides insights into the architecture of the deep learning model, indicating a well-structured neural network. The scoring history reflects consistent improvement over training epochs, affirming the model's learning capacity. This comprehensive analysis (Table 2) underscores the reliability and efficacy of the DNN in accurately predicting case-control outcomes.

Gene Ontology Analyses of DEGs, Functional Annotation, and Pathway Enrichment Analyses

To comprehend the biological activities, pathways, or locations of DEGs, gene ontology analyses concentrate on 3 domains—biological processes, cellular components, and molecular functions. Regarding biological processes, DEGs were primarily involved in tight voltage-gated potassium channels, transport of small molecules, transcriptional regulation of pluripotent stem cells, TP53-regulated transcription of genes involved in G1 cell cycle arrest, tight junction interactions, and PI3K/AKT signaling in cancer. The Reactome Pathway Analysis Database is a valuable resource for determining the potential roles that distinct DEGs may play in disease states and signaling pathways. For functional enrichment studies, the ClusterProfiler R package was used, with a significant cutoff set at $P < 0.05$ (Figure 2, A and B).

Investigation of the Correlations of Clinical/Demographic With Disease

The correlation matrix reveals associations among various demographic and clinical variables in the studied dataset. Weight demonstrates a moderately positive correlation with height ($r = 0.4$), suggesting that individuals with higher weights tend to be taller. Additionally, a small but statistically significant negative correlation is observed between weight and age ($r = -0.2$), indicating a tendency for younger individuals to have slightly higher weights. Notably, neoplasm cancer status exhibits a notable positive correlation with cancer stage ($r = 0.3$), suggesting that individuals with a positive cancer status are more likely to present with advanced stages of cancer. Moreover, a weak positive correlation is observed between race and ethnicity ($r = 0.1$), suggesting a tendency for certain racial and ethnic groups to share similar demographic characteristics. However,

Table 2. Selected Gene with Deep Learning Coefficient

| Gene | Variable Importance |
|----------------|---------------------|
| MEX3B 84206 | 1 |
| MYOZ3 91977 | 0.998401 |
| MBNL2 10150 | 0.988257 |
| VWF 7450 | 0.963862 |
| SPC25 57405 | 0.961849 |
| SBK1 388228 | 0.954074 |
| OLFM1 10439 | 0.947944 |
| ATRNL1 26033 | 0.946667 |
| TSPAN7 7102 | 0.946547 |
| MMRN1 22915 | 0.945444 |
| ECM2 1842 | 0.65634 |
| SPARCL1 8404 | 0.650928 |
| PRKAR2B 5577 | 0.649769 |
| PCDHB10 56126 | 0.647672 |
| GATA6 2627 | 0.646544 |
| MTUS2 23281 | 0.642114 |
| PAX8 7849 | 0.640089 |
| C1QTNF2 114898 | 0.632099 |
| DENND2A 27147 | 0.626046 |
| CLDN20 49861 | 0.619797 |

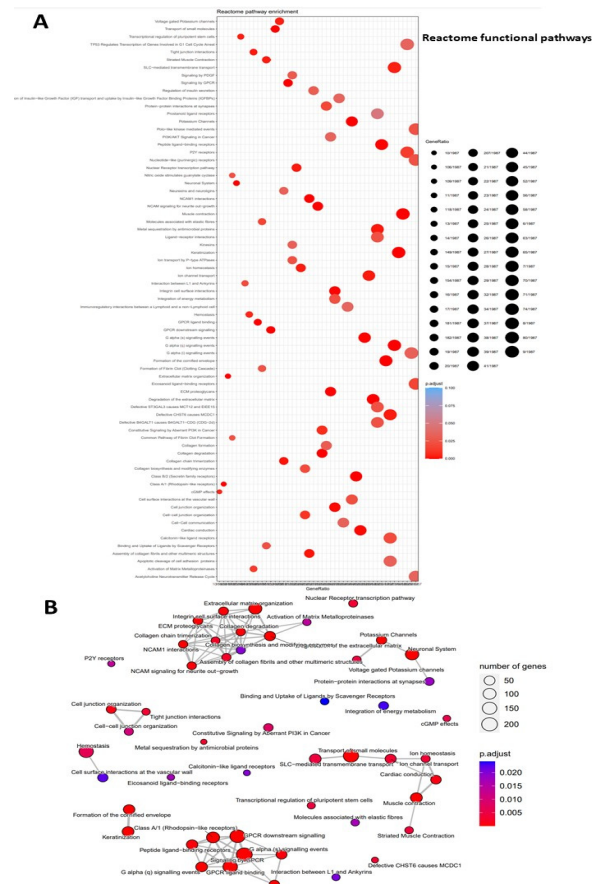


Figure 2. A, B. Reactome functional pathways

other correlations—such as those between age and ethnicity or age and case-control status—do not reach statistical significance (Figure 3B).

PPI Network Construction

The interactions among DEGs were analyzed and visualized using the STRING database, as illustrated in Figure 1B. An interaction score threshold of 0.4 was applied. The

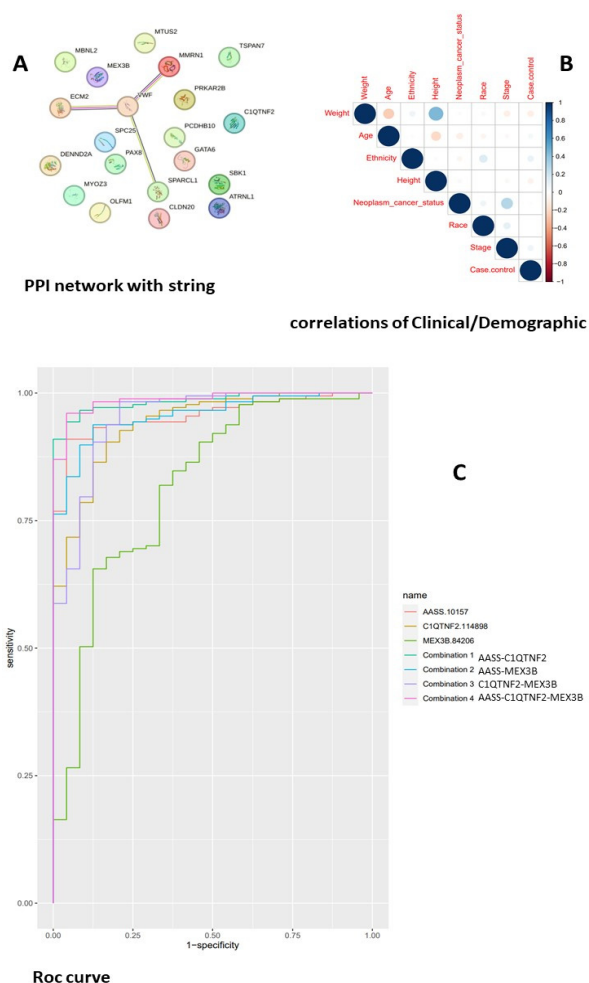


Figure 3. A PPI network of DEGs, B Correlation matrix shows significant co-relationship between clinical/demographic influence variables in UCEC, C combineROC curve of genes UCEC

analysis indicated a significant correlation among the VWF, ECM2, MMRN1, and SPARCL1 proteins (Figure 3A).

Identification of Diagnosis DEGs in UCEC

The ROC curve analysis was conducted to evaluate the diagnostic impact of key signature genes in UCEC. The results suggested that the genes MEX3B, CTRP2 (C1QTNF2), and AASS could serve as potential new biomarkers for UCEC patients. All analyses were carried out using SPSS Version 20, with $P < 0.05$ (Figure 3, A-C).

Genes Validation

The expression levels of candidate genes in UCEC patients were authenticated utilizing data from the GDAC available at <https://gdac.broadinstitute.org/>. GEO datasets—including GSE119041, GSE17025, GSE115810, GSE36389, and GSE25405—were employed for further verification. The validation dataset comprising data from patients with UCEC was obtained from this online resource, and subsequent preprocessing steps were executed to ensure data quality and consistency. The data showed that the mean expression of the candidate genes—

(MEX3B, CTRP2 (C1QTNF2), and AASS (mean \pm SD, 7.13 ± 5.6)—was higher in tumor cells ($P < 0.05$). Furthermore, there was no correlation between the dysregulation of the candidate genes—MEX3B, CTRP2 (C1QTNF2), AASS—and demographic and clinicopathological characteristics. As a result of checking other data sets, the candidate genes—MEX3B, CTRP2 (C1QTNF2), and AASS—were confirmed and valid.

Discussion

UCEC is the fourth most common cancer among women in the United States and the sixth most common cancer among women around the world (2). Finding novel biomarkers for prognosis prediction based on a possible molecular foundation of tumor development was our primary goal. The present study combined several bioinformatics and deep learning models to identify novel biomarkers of UCEC. MEX3B, an RNA-binding protein, is overexpressed in uterine endometrial cancer. Its upregulation has been associated with tumor growth, invasion, and metastasis (19). CTRP2, a member of the C1q/TNF-related protein family, is downregulated in uterine endometrial cancer. Its decreased expression has been correlated with poor prognosis and survival outcomes in patients with the disease (20). C1QTNF2, another member of the C1q/TNF-related protein family, has also been implicated in uterine endometrial cancer. Its dysregulation has been linked to tumor growth, angiogenesis, and metastasis (21). AASS, an enzyme involved in the biosynthesis of lysine and ketone bodies, is aberrantly expressed in uterine endometrial cancer. Its altered levels have been associated with metabolic reprogramming and tumor development. AASS can be used as a biomarker for monitoring metabolic changes in uterine endometrial cancer (22). The findings showed that the majority of important upregulated genes are connected to the tight voltage-gated potassium channels, transport of small molecules, transcriptional regulation of pluripotent stem cells, TP53 regulates transcription of genes involved in G1 cell cycle arrest, tight junction interactions, and PI3K/AKT signaling in cancer.

Almost all RNA posttranscriptional activities are regulated by RNA-binding proteins (RBPs), which are largely conserved across species and essential to maintaining gene expression homeostasis (23). Four members of the human MEX3 (muscle excess 3) family, which encodes diverse phosphorylated proteins and has distinct expression patterns, are members of the evolutionarily conserved RBP family (24). MEX3 is implicated in a variety of biological processes in the development and progression of cancer, which is consistent with the idea that cancer is a multi-pathway disease and the different functions that MEX3 plays in regulating gene expression (25). In multiple types of cancer—including bladder and breast cancer—MEX3 mediates migration, tumor immune escape mechanisms, cancer cell proliferation, and transcription level changes (26-29); depending on the type of tumor and MEX3 family member, MEX3 expression is correlated with either increased or decreased patient survival (30). MEX3B, or RNA binding family member B, is a translational regulator belonging to the MEX3 (muscle excess 3) family (24). To destabilize its

mRNA, MEX3B binds to the 3' long conserved untranslated region (3'UTR), which has components for both translational enhancement and mRNA destabilization (31). According to one study, MEX3B can function as TLR3's co-receptor during the innate antiviral response (32). By binding to the 3' UTR of HLA-A mRNA, MEX3B overexpression in melanoma cells can downregulate HLA-A expression, preventing T cells from identifying and eliminating tumor cells and causing resistance to immunotherapy (33). Furthermore, Mex3b was an E3 ligase that contributed to Runx3's widespread degradation brought on by HOTAIR. Runx3 degradation was reduced when HOTAIR or Mex3b expression was silenced. The expression level of Runx3 protein in human gastric cancer tissues was inversely correlated with HOTAIR (Pearson coefficient -0.501 ; $P = 0.025$). HOTAIR inhibition markedly reduced the migration and invasion of gastric cancer cells by upregulating claudin1, a process that could be reversed by co-deficiency in Runx3 (34). As demonstrated by our research, CTRP2 is a novel biomarker for UCEC. Adipokine superfamily member complement Cq1/tumor necrosis factor-related protein (CTRP), which is released from adipose tissue, shares a high degree of sequence similarity with adiponectin (35). According to earlier research, members of the CTRP family regulate a variety of physiological and pathological processes, such as the metabolism of carbohydrates and lipids, inflammation, the development and production of cartilage, cardiac protection, and vasodilation (36). Extensive research is being conducted on the involvement of the CTRP family in cancer. Currently, a number of the CTRP family members are considered to be molecular mediators that control the development of tumors as well as their invasion and metastasis (37). By activating several signal pathways, several CTRPs—including CTRP3, CTRP4, CTRP6, and CTRP8—have been reported to be related to osteosarcoma, hepatocellular cancer, colon cancer, and glioblastoma, respectively (38). Thus, in certain malignancies, CTRPs could act as both therapeutic targets and diagnostic markers. One of the most well-studied and useful adipocytokines that play a crucial role in regulating the body's metabolism is CTRP2, which shares 42% of the amino acid similarity with adiponectin at the functional globular C1q domain and is mostly expressed by adipose tissue (39). In plasma, this adipokine is also circulating as a trimer glycoprotein. According to earlier studies, CTRP2 regulates the metabolism of lipids and carbohydrates. Recombinant CTRP2 activates AMP-activated protein kinase (AMPK), a biological energy regulator, in muscle cells similar to adiponectin (40). The full-length and shortened forms of CTRP2 protein raise phosphorylation of AMPK, p42/44 MAPK, and acetyl-CoA carboxylase (ACC). Nevertheless, no prior research has examined the relationship between CTRP2 and malignancies (41). In line with previous studies showing a significant decrease in AASS expression in human breast cancer, overexpression of AASS or treatment with acetoacetate inhibited cell proliferation and induced autophagy and senescence in human cancer cell lines. Peroxisome proliferator-activated receptor γ (PPAR γ), a nuclear receptor that interacts with inflammatory mediators in obesity, is typically downregulated in human breast cancer. The expression of

2-aminoadipate semialdehyde synthase (AASS), which regulates the catabolism of lysine to acetoacetate, was found to be upregulated in the mammary epithelium of obese mice when PPAR γ expression was absent (42).

To find biomarkers in endometrial cancer, Wu et al used a novel bioinformatics approach that combines sample network building with GCN modeling. They selected features from non-Euclidean data using a graph convolutional network. Additionally, graph conventional networks, or GCNs, are specialists at assessing topologies with irregular structures, such as the interactions between chemicals and diseases. A total of 23 potential biomarkers were found. After conducting functional analyses to rationalize these biomarkers, network entropy characterization revealed a correlation between the biomarkers and illness survival. Future studies looking at the molecular causes and potential treatment targets of endometrial malignancies will benefit from these biomarkers (43). Another study identified several potential biomarker genes through transcriptomics and methylomics data analysis in patients with endometriosis. From transcriptomics, the candidate genes included NOTCH3, SNAPC2, B4GALNT1, SMAP2, DDB2, GTF3C5, and PTOV1, while from methylomics, the genes TRPM6, RASSF2, TNIP2, RP3-522J7.6, FGD3, and MFSD14B were highlighted. The study found that TMM normalization for transcriptomics data, quantile or Voom normalization for methylomics data, GLM for feature space reduction, and techniques to maximize classification performance should be incorporated into an effective machine learning diagnostic process for endometriosis (44). Furthermore, the dysregulation of biomarkers—eg, PKM, RAN, PHGDH, and SLC7A5—was linked to poorer survival rates in endometrial cancer patients. Suman et al employed machine learning classifiers—including Principal Component Analysis, Random Forest, Multinomial Naïve Bayes, and Support Vector Machine with Recursive Feature Elimination—to assess which interacting DEGs were most significant. Key overlapping DEGs, hub proteins, and important modules from PPI network analysis were proposed as candidate biomarkers for progressive endometrial cancer (45).

Conclusion

The GDCA database and deep learning algorithms identified 3 significant genes as potential diagnosis biomarkers of UCEC. Thus, identifying new UCEC biomarkers has promise for effective care, improved prognosis, and early diagnosis. The biomarkers MEX3B, CTRP2, C1QTNF2, and AASS play a critical role in understanding the diagnosis of uterine endometrial cancer by providing insights into the molecular mechanisms underlying the disease and facilitating early detection, prognosis, and treatment strategies. It is recommended that further research be conducted in the field of diagnostic genes for various diseases by employing a combination of artificial intelligence and bioinformatics analysis. Given the significant potential of this approach, additional studies in this area are warranted to enhance our understanding of disease pathology and improve diagnostic accuracy.

Authors' Contributions

Study conception and design: E.N., H.V.L., M.R., K.A., and A.M. Acquisition of data: E.N. Analysis and interpretation of data: E.N., A.M., H.V.L. Drafting of the manuscript: H.V.L. Critical revision: E.N.

Ethical Considerations

The Student Research Committee of Shahid Beheshti University of Medical Sciences approved this study with the code 43010456.

Acknowledgment

The authors acknowledge the financial support from Shahid Beheshti University of Medical Sciences.

Conflict of Interests

The authors declare that they have no competing interests.

References

- Besso MJ, Montivero L, Lacunza E, Argibay MC, Abba M, Furlong LI, et al. Identification of early stage recurrence endometrial cancer biomarkers using bioinformatics tools. *Oncol Rep.* 2020;44(3):873-86.
- Geng R, Zheng Y, Zhou D, Li Q, Li R, Guo X. ZBTB7A, a potential biomarker for prognosis and immune infiltrates, inhibits progression of endometrial cancer based on bioinformatics analysis and experiments. *Cancer Cell Int.* 2020;20(1):542.
- Henley SJ, Miller JW, Dowling NF, Benard VB, Richardson LC. Uterine Cancer Incidence and Mortality - United States, 1999-2016. *MMWR Morb Mortal Wkly Rep.* 2018;67(48):1333-8.
- Uterine Cancer - Statistics. *Cancer.Net.* June 25, 2012. [Available from: <https://www.cancer.net/cancer-types/uterine-cancer/statistics>.
- Dai F, Chen G, Wang Y, Zhang L, Long Y, Yuan M, et al. Identification of candidate biomarkers correlated with the diagnosis and prognosis of cervical cancer via integrated bioinformatics analysis. *Onco Targets Ther.* 2019;12:4517-32.
- Nazari E, Chang H-CH, Deldar K, Pour R, Avan A, Tara M, et al. A comprehensive overview of decision fusion technique in healthcare: A systematic scoping review. *Iranian Red Crescent Medical Journal.* 2020;22(10).
- Song Y, Chen QT, He QQ. Identification of key transcription factors in endometrial cancer by systems bioinformatics analysis. *J Cell Biochem.* 2019;120(9):15443-54.
- Huang CY, Liao KW, Chou CH, Shrestha S, Yang CD, Chiew MY, et al. Pilot Study to Establish a Novel Five-Gene Biomarker Panel for Predicting Lymph Node Metastasis in Patients With Early Stage Endometrial Cancer. *Front Oncol.* 2019;9:1508.
- Nazari E, Pourali G, Khazaei M, Asadnia A, Dashtihangar M, Mohit R, et al. Identification of Potential Biomarkers in Stomach Adenocarcinoma using Machine Learning Approaches. *Current Bioinformatics.* 2023;18(4):320-33.
- Zhao H, Jiang A, Yu M, Bao H. Identification of biomarkers correlated with diagnosis and prognosis of endometrial cancer using bioinformatics analysis. *J Cell Biochem.* 2020;121(12):4908-21.
- Ye Y, Li H, Bian J, Wang L, Wang Y, Huang H. Exploring Prognosis-Associated Biomarkers of Estrogen-Independent Uterine Corpus Endometrial Carcinoma by Bioinformatics Analysis. *Int J Gen Med.* 2021;14:9067-81.
- Zagidullin B, Pasanen A, Loukovaara M, Bützow R, Tang J. Interpretable prognostic modeling of endometrial cancer. *Sci Rep.* 2022;12(1):21543.
- Zhao H, Jiang A, Yu M, Bao H. Identification of biomarkers correlated with diagnosis and prognosis of endometrial cancer using bioinformatics analysis. *J Cell Biochem.* 2020;121(12):4908-21.
- Li Y, Li L. Bioinformatic screening for candidate biomarkers and their prognostic values in endometrial cancer. *BMC Genet.* 2020;21(1):113.
- Bhatt, Chandradeep, et al. The state of the art of deep learning models in medical science and their challenges. *Multimedia Systems.* 2021;27(4):599-613.
- Aggarwal, Charu C. *Neural networks and deep learning.* Springer. 2018;10(978):3.
- Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv.* 2016;1605.07678.
- Fergus P, Chalmers C. *Performance Evaluation Metrics.* Applied Deep Learning: Springer; 2022. p. 115-38.
- Yang, Yan, et al. The RNA-binding protein Mex3B is a coreceptor of Toll-like receptor 3 in innate antiviral response. *Cell Res.* 2016;26(3):288-303.
- Jerobin, Jayakumar, et al. Regulation of circulating CTRP-2/CTRP-9 and GDF-8/GDF-15 by intralipids and insulin in healthy control and polycystic ovary syndrome women following chronic exercise training. *Lipids Health Dis.* 2021;20:1-11.
- Lei, Xia, and G. William Wong. C1q/TNF-related protein 2 (CTRP2) deletion promotes adipose tissue lipolysis and hepatic triglyceride secretion. *J Biol Chem.* 2019;294(43):15638-15649.
- Naseer, Muhammad Imran, et al. Next generation sequencing reveals novel homozygous frameshift in PUS7 and splice acceptor variants in AASS gene leading to intellectual disability, developmental delay, dysmorphic feature and microcephaly. *Saudi J Biol Sci.* 2020;27(11):3125-3131.
- Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol.* 2018;19(5):327-41.
- Buchet-Poyau K, Courchet J, Hir HL, Séraphin B, Scoazec JY, Duret L, et al. Identification and characterization of human Mex-3 proteins, a novel family of evolutionarily conserved RNA-binding proteins differentially localized to processing bodies. *Nucleic Acids Res.* 2007;35(4):1289-300.
- Bufalieri F, Caimano M, Lospinoso Severini L, Basili I, Paglia F, Sampirisi L, et al. The RNA-binding ubiquitin ligase MEX3A affects glioblastoma tumorigenesis by inducing ubiquitylation and degradation of RIG-I. *Cancers.* 2020;12(2):321.
- Hu B, Yang XB, Sang XT. Development and verification of the hypoxia-related and immune-associated prognosis signature for hepatocellular carcinoma. *J Hepatocell Carcinoma.* 2020;3:15-30.
- Jia R, Weng Y, Li Z, Liang W, Ji Y, Liang Y, et al. Bioinformatics analysis identifies IL6ST as a potential tumor suppressor gene for triple-negative breast cancer. *Reprod Sci.* 2021;28:2331-41.
- Huang Y, Fang C, Shi JW, Wen Y, Liu D. Identification of hMex-3A and its effect on human bladder cancer cell proliferation. *Oncotarget.* 2017;8(37):61215.
- Zhang J, Zhou YJ, Yu ZH, Chen AX, Yu Y, Wang X, et al. Identification of core genes and clinical roles in pregnancy-associated breast cancer based on integrated analysis of different microarray profile datasets. *Biosci Rep.* 2019;39(6):BSR20190019.
- Tian M, Qiao M, Shen C, Meng F, Frank LA, Krasitskaya VV, et al. Highly-sensitive graphene field effect transistor biosensor using PNA and DNA probes for RNA detection. *Applied Surface Science.* 2020;527:146839.
- Takada H, Kawana T, Ito Y, Kikuno RF, Mamada H, Araki T, et al. The RNA-binding protein Mex3b has a fine-tuning system for mRNA regulation in early *Xenopus* development. *Journal abbrev.* 2009.
- Yang Y, Wang SY, Huang ZF, Zou HM, Yan BR, Luo WW, et al. The RNA-binding protein Mex3B is a coreceptor of Toll-like receptor 3 in innate antiviral response. *Cell Res.* 2016;26(3):288-303.
- Huang L, Malu S, McKenzie JA, Andrews MC, Talukder AH, Tieu T, et al. The RNA-binding protein MEX3B mediates resistance to cancer immunotherapy by downregulating HLA-A expression. *Clin Cancer Res.* 2018;24(14):3366-76.
- Xue M, Chen LY, Wang WJ, Su TT, Shi LH, Wang L, et al. HOTAIR induces the ubiquitination of Runx3 by interacting with Mex3b and enhances the invasion of gastric cancer cells. *Gastric Cancer.* 2018;21:756-64.
- Wong GW, Krawczyk SA, Kitidis-Mitrokostas C, Revett T, Gimeno R, Lodish HF. Molecular, biochemical and functional characterizations of C1q/TNF family members: adipose-tissue-selective expression patterns, regulation by PPAR-γ agonist, cysteine-mediated oligomerizations, combinatorial associations and metabolic functions. *Biochem J.* 2008;416(2):161-77.
- Server DC, Stewart AN, Rodriguez S, Little HC, Aja S, Wong GW. Loss of CTRP4 alters adiposity and food intake behaviors in obese mice. *American Journal of Physiology-Endocrinology and Metabolism.* 2020;319(6):E1084-E1100.
- Kong M, Gao Y, Guo X, Xie Y, Yu Y. Role of the CTRP family in tumor development and progression. *Oncol Lett.* 2021;22(4):1-9.

38. Lv W, Li Q, Jia B, He Y, Ru Y, Guo Q, et al. Differentiated embryonic chondrocyte-expressed gene 1 promotes temozolomide resistance by modulating the SP1-MGMT axis in glioblastoma. *American Journal of Translational Research*. 2021;13(4):2331.
39. Lui H, Gao X, Zhou Y, Li N, Ge C, Hui X, et al. High level expression, purification and characterization of active fusion human C1q and tumor necrosis factor related protein 2 (hCTRP2) in *Escherichia coli*. *Protein Expression and Purification*. 2011;79(1):1-6.
40. Ilbeigi D, Khoshfetrat M, Afrisham R, Rahimi B, Gorgani-Firuzjaee S. Serum C1q/TNF-related Protein-2 (CTRP2) levels are associated with coronary artery disease. *Archives of Medical Research*. 2020;51(2):167-72.
41. Wong GW, Wang J, Hug C, Tsao T-, Lodish HF. A family of Acrp30/adiponectin structural and functional paralogs. *Proceedings of the National Academy of Sciences*. 2004;101(28):10302-7.
42. Wu J, Kramer K, Crowe DL. Lysine metabolism is a novel metabolic tumor suppressor pathway in breast cancer. *Oncogene*. 2023;42(31):2402-14.
43. Wu E, Fan X, Tang T, Li J, Wang J, Liu X, et al. Biomarkers discovery for endometrial cancer: A graph convolutional sample network method. *Computers in Biology and Medicine*. 2022;150:106200.
44. Akter S, Xu D, Nagel SC, Bromfield JJ, Pelch K, Wilshire GB, et al. Machine learning classifiers for endometriosis using transcriptomics and methylomics data. *Frontiers in genetics*. 2019;10:466838.
45. Suman S, Kulshrestha A. Integrative approach: Advancing endometrial cancer biomarkers through machine learning and network-derived RNA-Seq screening. *Human Gene*. 2024;39:201257.