scientific reports

Check for updates

OPEN

Systems biology and machine learning approaches identify drug targets in diabetic nephropathy

Maryam Abedi[1], Hamid Reza Marateb[2,3], Mohammad Reza Mohebian[4], Seyed Hamid Aghaee-Bakhtiari[5,6], Seyed Mahdi Nassiri[7] & Yousof Gheisari[1,8]✉

Diabetic nephropathy (DN), the leading cause of end-stage renal disease, has become a massive global health burden. Despite considerable efforts, the underlying mechanisms have not yet been comprehensively understood. In this study, a systematic approach was utilized to identify the microRNA signature in DN and to introduce novel drug targets (DTs) in DN. Using microarray profiling followed by qPCR confirmation, 13 and 6 differentially expressed (DE) microRNAs were identified in the kidney cortex and medulla, respectively. The microRNA-target interaction networks for each anatomical compartment were constructed and central nodes were identified. Moreover, enrichment analysis was performed to identify key signaling pathways. To develop a strategy for DT prediction, the human proteome was annotated with 65 biochemical characteristics and 23 network topology parameters. Furthermore, all proteins targeted by at least one FDA-approved drug were identified. Next, *mGMDH-AFS*, a high-performance machine learning algorithm capable of tolerating massive imbalanced size of the classes, was developed to classify DT and non-DT proteins. The sensitivity, specificity, accuracy, and precision of the proposed method were 90%, 86%, 88%, and 89%, respectively. Moreover, it significantly outperformed the state-of-the-art (*P*-value ≤ 0.05) and showed very good diagnostic accuracy and high agreement between predicted and observed class labels. The cortex and medulla networks were then analyzed with this validated machine to identify potential DTs. Among the high-rank DT candidates are Egfr, Prkce, clic5, Kit, and Agtr1a which is a current well-known target in DN. In conclusion, a combination of experimental and computational approaches was exploited to provide a holistic insight into the disorder for introducing novel therapeutic targets.

Diabetic nephropathy (DN) is a major complication in diabetes mellitus and the leading cause of end-stage renal disease (ESRD). Despite the beneficial effects of current drugs such as angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers, DN patients are still reaching ESRD[1]. Therefore, it is critical to understand the molecular mechanisms of this disorder to develop more efficient therapeutic approaches. Systems biology with its quantitative and predictive viewpoints has provided a unique opportunity to explore the complex biological processes involved in the pathogenesis of chronic disorders[2,3]. It allows the generation of holistic maps of interactions between a variety of biomolecules involved in these processes. Considering the pivotal role of microRNAs (miRNAs) in the regulation of a bundle of functionally related genes[4,5], we were motivated to study the network of DN-associated miRNAs with their targets. Although the role of individual miRNAs in DN has been previously explored[6], holistic evaluations have just been started[7].

Valuable insights into the molecular mechanisms of complex disorders have been introduced using systems biology strategies, however, little progress has been made in the translation of this knowledge to the bedside. Although omics technologies, in line with advanced computational techniques, allowed the identification of lots of biomolecules with complex interactions in disease pathogenesis, the identification of appropriate therapeutic

[1]Regenerative Medicine Research Center, Isfahan University of Medical Sciences, Isfahan, Iran. [2]Biomedical Engineering Department, Engineering Faculty, University of Isfahan, Isfahan, Iran. [3]Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya, BarcelonaTech (UPC), Barcelona, Spain. [4]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, Canada. [5]Bioinformatics Research Group, Mashhad University of Medical Sciences, Mashhad, Iran. [6]Department of Medical Biotechnology and Nanotechnology, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. [7]Department of Clinical Pathology, Faculty of Veterinary Medicine, University of Tehran, Tehran, Iran. [8]Department of Genetics and Molecular Biology, Isfahan University of Medical Sciences, Isfahan, Iran. ✉email: ygheisari@med.mui.ac.ir

targets has remained elusive. To address this challenge, some investigators have proposed that the success of current FDA-approved drugs compared to many ingredients that failed during preclinical and clinical screenings can, at least partly, be attributed to the characteristics of their target proteins[8]. Hence, these potential discriminating properties can be identified and exploited to predict novel drug targets (DT). Based on this assumption, several classic machine learning algorithms have been utilized for feature selection and DT prediction[9–15]. However, these studies suffer from several limitations such as ignoring the unequal frequency of DT and non-DT proteins, inappropriate machine performance measures, or unspecified details of utilized methods. We have here developed a next-generation machine learning method that considers high-level feature interactions and unbalanced DT/non-DT classes.

This study aimed to predict novel targets for DN based on the holistic map of molecular pathogenesis. Several in silico and wet lab steps were pursued to identify the miRNA profile of the disorder and detect novel differentially expressed (DE) miRNAs in the cortex and medulla of diabetic kidneys. Moreover, miRNA-target interaction networks were inferred to identify central nodes and critical interactions. Pathway enrichment analysis also allowed the prediction of affected signaling pathways in this disorder. Next, to translate the findings of this study to clinical application, a high-performance machine learning framework, named "modified Group Method of Data Handling with Automatic Feature Selection (mGMDH-AFS)", was developed and validated for the prediction of DT in human proteome based on a variety of biochemical and network topology features. This classifier was then applied to candidate novel therapeutic targets in the constructed holistic map of DN. The design of this study is schematically presented in Fig. 1.

Materials and methods

Diabetic nephropathy mouse model.

Male DBA/2 J mice, aged 6–10 weeks, were supplied from the Pasteur Institute of Iran (Tehran, Iran). All animal studies were performed according to the NIH guide for the care and use of laboratory animals[16]. In addition, the study was conducted in compliance with the ARRIVE guidelines. All protocols were approved by the Isfahan University of Medical Sciences Ethics Committee (IR.MUI.MED.REC.1399.933). For five consecutive days, streptozotocin (STZ) was dissolved in sodium citrate buffer and intraperitoneally administered to pre-starved mice at the dose of 40 mg/kg (total amount: 200 mg/kg). Control mice received citrate buffer. Mice were supplied with 10% sucrose water during STZ injection and a few days after to avoid sudden hypoglycemia. As the mice with a lesser extent of diabetes usually do not show renal injury, one week after the final STZ injection, their non-fasting blood glucose was evaluated and those with blood glucose levels below 280 mg/dL were excluded from the study. Three months after the last dose of STZ, mice with the blood glucose range of 300–600 mg/dL were transferred to metabolic cages to collect 24-h urine. Urine albumin concentration was measured with an ELISA test (Exocell, Philadelphia, PA), and urine volume was used to calculate total albumin excretion. Urine-specific gravity, kidney weight, serum glucose, HbA1C, serum creatinine, and urea were also measured. Moreover, kidney tissues were sampled. Histopathologic parameters such as glomerular basement membrane thickening, mesangial matrix, increased mesangial cell proliferation, and diffuse mesangial sclerosis were assessed after Hematoxylin and Eosin (H&E) and Periodic Acid–Schiff (PAS) staining. For each pathology field, 20 serial glomeruli were evaluated, and the percentages of affected glomeruli were calculated. The Mann–Whitney U test was applied for statistical analysis.

miRNA microarray.

The left kidneys were harvested, and the cortex and medulla were separated. The tissues were kept in RNAlater (Qiagen, Valencia, CA, USA) and stored at −70 °C until RNA extraction. The tissues were lysed in 1 mL QIAzol (Qiagen, Valencia, CA, USA) and homogenized with TissueLyser LT (Qiagen). Next, 250 μL chloroform (Merck, Darmstadt, Germany) was added and centrifuged at 12,000 rpm for 20 min at 4 °C after 15 min of incubation. An equal volume of cold ethanol (Merck) was added to the upper aqueous phase in a new tube and incubated at −20 °C overnight after which, the samples were centrifuged at 14,000 g for 45 min at 4 °C. To dislodge the pellet, 75% ethanol was added and centrifuged at 12,000 g for 15 min at 4 °C. The dried pellet was dissolved in double-distilled water. Using a BioPhotometer (Eppendorf, Hamburg, Germany), RNA concentration was measured at 260 nm.

miRNA profiling was performed for the cortex and medulla of five DN and three control mice (total 16 samples). The quality of the microarray experiment was evaluated by the principal component analysis (PCA) and the hierarchical clustering. For hierarchical clustering, the correlation coefficient, and the average linkage methods, as the distance metric, were applied and the heat maps were plotted by ClusterMaker application[17] of Cytoscape software version 3.2.0[18]. PCA was carried out with ggplot package[19] of R[20]. The miRNAs with logarithm to base two of fold change ($\log_2 FC$) ≥ 0.5 or ≤ -0.5 were selected.

Quantitative PCR.

Primers were designed with Gene Runner version 3.05 (Hastings Software Inc., Hastings, NY, USA) and Oligo version 7 (Molecular Biology Insights, Inc, USA) software. cDNA was synthesized using RevertAid First-Strand cDNA Synthesis Kit (Thermo Scientific, Vilnius, Lithuania) and PCR thermal cycler machine (Takara Bio, Shiga, Japan). The RNA samples, stem-loop specific primer, and double-distilled water were mixed and incubated at 75 °C for 5 min. The vials were immediately placed on ice, and ten mM dNTP mix, 5X reaction buffer, M-MuLV RT enzyme, and double-distilled water were added and spun briefly. Next, they were incubated at 42 °C for 60 min. The reaction was terminated by heating at 70 °C for 5 min.

QuantiFast SYBR Green PCR kit (Takara Bio) and Rotor-Gene 6000 real-time PCR machine (Corbett, Sydney, Australia) were employed to assess the expression levels of 37 miRNAs as well as Snord70 and Snord68, as internal controls in the same samples used for the microarray experiment. The temperature profile consisted of an initial step of 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. The data were analyzed using REST 2009 software[21].

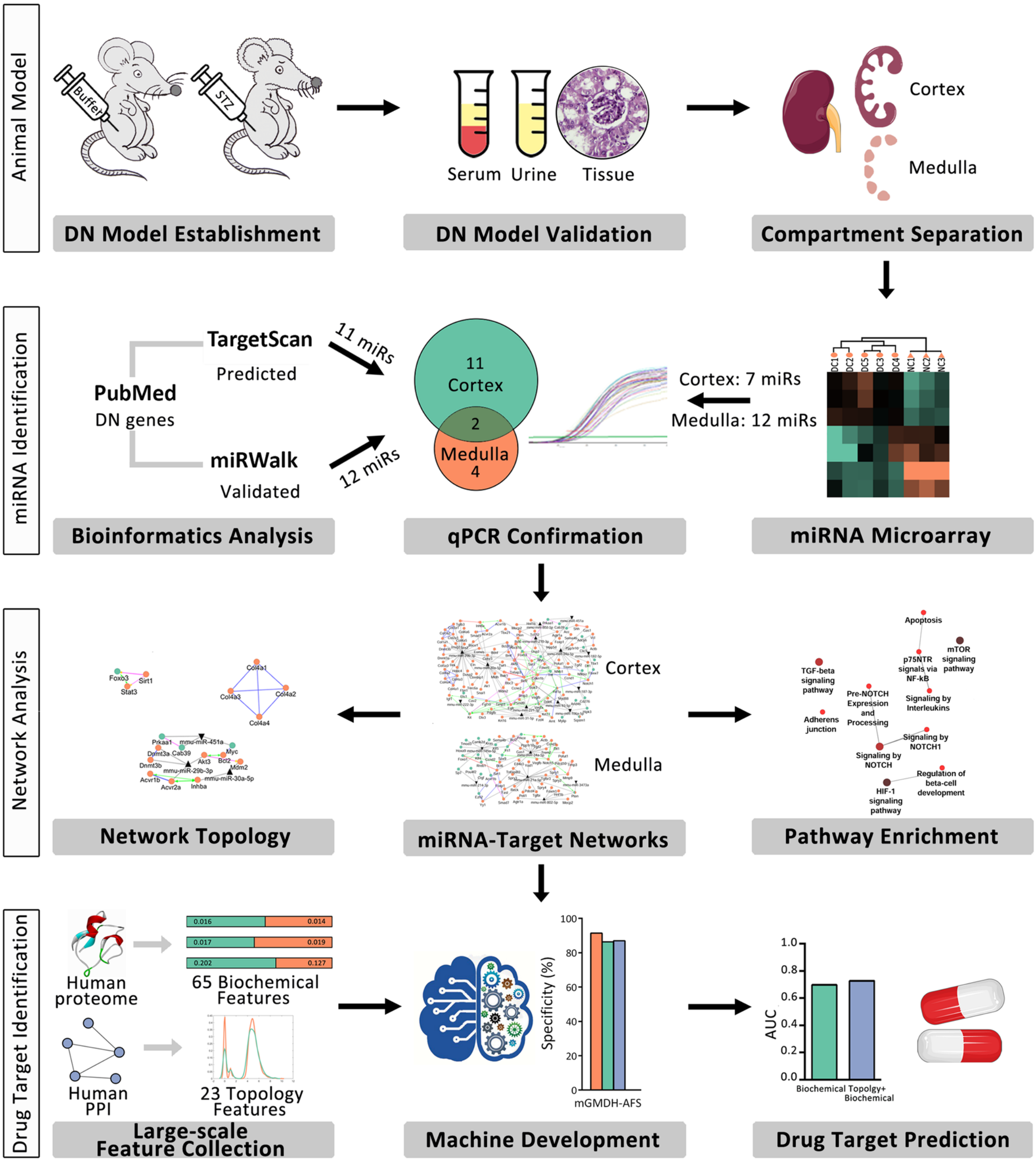

Figure 1. A schematic representation of the study design. This study was aimed at predicting novel drug targets for DN based on the holistic molecular pathogenesis map. Using different experimental and computational methods, the central nodes, key interactions, and signaling pathways of DN were identified. To translate the findings to clinical application, a high-performance machine learning framework, *mGMDH-AFS,* was developed and validated to predict drug targets for all human proteins. This classifier was then applied to candidate novel therapeutic targets in the constructed holistic map of DN. miRs: microRNAs; PPI: protein–protein interaction network.

MicroRNA target identification. Experimentally validated targets of DE miRNAs were collected using miRTarBase database[22]. Targets that were validated by strong evidence were selected. The validated targets were obtained from studies on murine samples for all miRNAs, except for mmu-miR-802-5p and mmu-miR-187-3p that did not have any mouse validated targets; thus, human targets with strong evidence were obtained. A list of genes with potential roles in DN was manually collected from the literature. TargetScan 6.2[23] and miRWalk databases[24] were used for predicted, and validated miRNAs targeting these genes, respectively. The miRNAs that were broadly conserved among vertebrates and mammals were selected in TargetScan.

Network construction and analysis. Using CluePedia plugin version 2.1.7[25] of Cytoscape software, interaction networks for validated targets of DE miRNAs in the cortex and medulla were constructed by STRING resource[26]. Confidence cut-off was set at 0.8; then, DE miRNAs were merged to the networks. The edges including activation, post-translational modification, binding, and inhibition were allowed to be shown. The topology of the networks was analyzed by NetworkAnalyzer[27], CytoNCA[28], and CentiScape[29] applications of Cytoscape. Mann–Whitney U test was employed to compare the topology features using the SPSS statistical package, version 18.0 (SPSS Inc., Chicago, IL, USA).

The modules were determined in the networks by the MCODE version 1.4.1[30] plugin of Cytoscape. Human protein–protein interactions were retrieved from HPRD version 9[31] and imported to the Cytoscape.

Gene ontology and pathway enrichment analysis.

Pathway enrichment analysis was carried out using Cytoscape ClueGO plugin version 2.1.7[32]. In this analysis, Reactome[33] and KEGG[34] databases were selected for retrieving the data. In addition, the DAVID database[35] was applied to collect pathways and biological processes from BioCarta and Gene Ontology (GO), respectively. P-value correction was performed and pathways and GO terms with the adjusted P-value ≤ 0.05 were selected.

Drug target collection.

FDA-approved drugs and their target proteins were downloaded from the DrugBank database[36]. Non-human targets were deleted.

Collection of biochemical features.

Protein accession and annotation. The UniProt ID and the official protein name for all human proteins were extracted from the UniProtKB/Swiss-Prot of the UniProt database[37]. Using SPSS statistical package, the Chi-square test was employed to compare the biochemical features.

Protein family. Human proteins belonging to receptors, G-Couple Protein Receptors (GPCRs), nuclear hormone receptors, enzyme-linked receptors, tyrosine kinase receptors, serine/threonine receptors, ion channels, ligand, and voltage-gated ion channels, transporters, GTPases, ATPases, Phosphatase, and proteases were extracted from protein family of UniProt. In addition, human kinases were collected from the pkinfam data of UniProt.

Post-translational modification. Experimentally validated post-translational modifications (PTM) were extracted from the dbPTM database[38].

Enzyme. Enzymes were downloaded from the ENZYME database[39] and categorized into oxidoreductase, transferase, hydrolase, lyase, isomerase, and ligase classes. Metabolic enzymes were extracted from Metabolic Enzyme Database[40].

Transcription factor and cofactor. The experimentally validated human transcription factors (TFs) list was extracted from the TFCheckpoint database[41]. Transcription cofactors were collected from AnimalTFDB[42].

Epigenetics regulator. The epigenetics regulators including chromatin remodelers, histone modifiers, RNA/DNA modifiers, and scaffold proteins were extracted from EpiFactors[43].

Transcriptional response to small molecules. The list of up or down-regulated genes in the treatment of small molecules identified in the CMAP project was harvested from Enrichr[44].

miRNA target. All human genes identified as strongly validated miRNA targets were retrieved from miRTarBase[22].

Mitochondrial protein. Human proteins located in mitochondria were collected from miToCarta2.0[45].

Mutation. A list of human mutated genes was downloaded from Online Mendelian Inheritance in Man (OMIM)[46].

SNPs-trait association. All SNP-trait associations with P-value ≤ 5×10^{-8} were obtained from GWAS catalog v2.2.1[47]. SNPs mapped in non-coding and intragenic regions were deleted. Coding genes that were nearest to the SNPs were collected.

Machine learning.

State-of-the-art. Current machine learning methods including logistic regression (LR)[14,48,49], radial basis function (RBF) kernel support vector machine (SVM)[9,12], generalized linear model (GLM)[50], and radial basis function network (RBFN)[14] were utilized to predict potential drug targets. Discriminative features were selected in LR and GLM based on their statistical structures. Sequential Forward Selection (SFS) method was used in SVM and RBFN[51]. The classifiers are briefly described below:

(a) Logistic regression

LR uses the following regression model for the prediction:

$$\log\left(\frac{p}{1-p}\right) = b0 + \sum_{j=1}^{N_f} b_j x_j + e$$

where p is the detection probability, and e denotes the binomial error term. Having fitted the model[52] by tuning the parameters b_j using the input features x_j (N_f is the number of features), each case with the estimated $p \geq 0.5$ was classified as the DT, or non-DT otherwise. Upon normalization of the input features, those with low weights could be excluded (a.k.a., feature selection).

(b) Support vector machine

Hyperplanes are mainly used in SVM to separate data points of different classes[51]. The data is transferred to higher dimensions using non-linear mappings (i.e., kernels). We used radial basis function kernels in this study. The method proposed by Wu and Wang was used to tune the radius of the RBF and the soft-margin parameter[53]. Moreover, the SVM classifier was trained using sequential minimal optimization[54].

(c) Generalized linear model

GLM is a flexible linear regression that involves using a link function to relate the output of the model to the response variable. LR is one of the categories of GLM which employs the logit function i.e., log (p/(1−p)) as the linking function. GLM is suitable for the binomial model. Here, the Poisson linking function was utilized[55,56].

(d) Radial basis function neural networks

RBF Networks encompass the following layers, the input as the entire features, a hidden layer with a non-linear RBF activation function, and a linear output layer with one node per category or class of data. Each output node calculates the score for the associated class and the class with the highest score is selected for each input sample. The RBF prototypes were estimated using the K-means clustering while the other network parameters were estimated using the Backpropagation algorithm[57].

The mGMDH-AFS algorithm. An mGMDH-AFS algorithm was developed based on inductive neural networks or Group Method of Data Handling (GMDH) that created more accessible models and provide more transparency[58,59]. In biological systems characterized by high dimensionality, it is crucial to perform feature selection to improve classification accuracy[60,61]. The GMDH algorithm was embedded with Particle Swarm Optimization (PSO), a population-based stochastic optimization algorithm[62], and a Relief-feature-weighting algorithm[63] to estimate optimal model fitting.

Briefly, categorical data were transferred to interval features using logistic regression[64,65]. The binary encoding was used for each categorical variable, and the logistic regression function parameters were estimated using iterative reweighted least-squares[66] on the estimation set. It is, indeed, a form of non-linear data normalization. The I-RELIEF algorithm was employed prior to the classification. The weight of the features was iteratively estimated based on their capability to discriminate between neighboring patterns[67]. Moreover, the oversampled[68] training set was divided into estimation and validation sets to avoid over-fitting[69].

The GMDH, proposed by Ivakhnenko[70] was utilized in the current study. In this network, the pairwise interactions of each input feature (a.k.a., neurons) are calculated at each layer. The output of each neuron in the current layer is used as the input to the next layer, and the network is built layer by layer until no improvement is observed in the validation set (i.e., early stopping criterion). The overall structure of the algorithm is provided in Fig. S1.

In this study, the top 10 neurons were selected at each layer. Instead of the traditional polynomial function widely used in the GMDH network, a matrix of nonlinear non-convex functions including exponential, sinusoid, and logarithmic forms was used to model the interaction. Thus, the PSO algorithm was used to estimate the parameters of the model rather than the least-square algorithm. Since the data is highly imbalanced, the Matthews correlation coefficient was used as the fitness function[71] instead of the traditional RMSE.

The output of the GMDH algorithm is a continuous variable ranging from zero to one. The optimal cut-off was then estimated on the training set using Receiver Operating Characteristic (ROC) curve[72]. The MATLAB code is available online to interested readers: https://github.com/marateb/Drug-Targets-Classification.

Validation. The hold-out validation was used to assess the performance of the developed methods. The dataset was randomly split into two independent 70% training and 30% test sets[73]. Also, four-fold cross-validation was employed for further performance assessment to guard against testing hypotheses suggested by the data (Type III errors)[74]. The classifiers were assessed in terms of the performance indices such as sensitivity, specificity, precision, accuracy, and diagnostic odds ratio (DOR) whose definitions are given in Table S1. The Q-Cochran's test and McNemar's post-hoc test were used to identify whether the proposed system significantly outperformed the state-of-the-art. The Bonferroni correction was also applied for multiple comparisons and the adjusted P-values were then used for interpretation. The random permutation test[75] was used to compare the performance of the mGMDH-AFS machine with real DT/non-DT classes and ten random sets whose class labels were randomly permuted. MATLAB version 8.6 (The MathWorks Inc., Natick, MA, USA) was used for offline processing.

Results

A combination of computational and experimental methods was employed to identify the miRNA profile in DN.

To explore DN pathogenesis, a mouse model of STZ-induced DN was established and after three months was validated using different functional (Fig. 2a) and histopathological (Fig. 2b-e) assessments. For constructing a holistic map of DN, we started with the profile of miRNA related to this disease as these molecules target functionally related genes so each variably expressed miRNA can be a clue to identify a group of related altered genes and functions[5, 76]. miRNA microarray was performed on the cortex and medulla samples, separately, and the quality of microarray data was confirmed using unsupervised hierarchical clustering and PCA (Fig. 3a). We identified 7 and 12 miRNAs with |logFC|≥0.5 in cortex and medulla, respectively (Fig. 3b). To propose further miRNAs, a list of genes with a documented role in DN was provided (Table S2). Using TargetScan and miRWalk databases, the predicted and validated miRNAs targeting the DN-associated genes were chosen, respectively (Table S3). Based on the microarray, miRNA prediction, and validation, a total

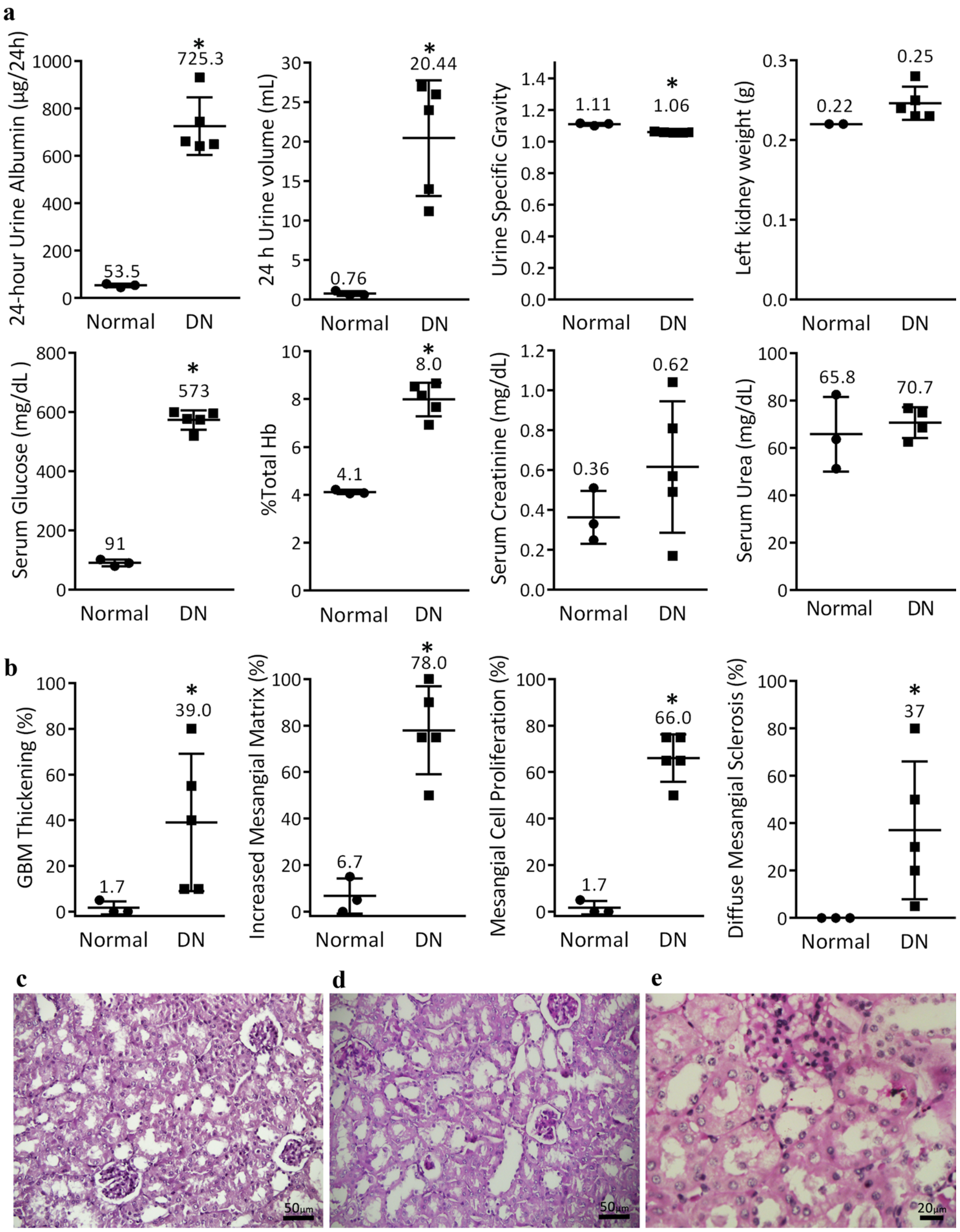

Figure 2. The mouse model of DN was validated with different parameters. A mouse model of DN was established using streptozotocin and validated after 3 months using functional (**a**) and histopathological (**b**) assessments. Representative fields of normal (**c**) and DN (**d**, **e**) kidneys are shown. Data are reported as means ± SD. Asterisks represent *P*-value ≤ 0.05. GBM: glomerular basement membrane.

of 37 miRNAs were considered to be potentially related to DN (Fig. 3c). The alternations in the expression of these candidate miRNAs were examined in cortex and medulla samples by quantitative PCR (qPCR). Despite several optimizations, a reliable quantification was not achieved for 8 miRNAs (mmu-miR-711, mmu-miR-592-3p, mmu-miR-186-5p, mmu-miR-495-3p, mmu-miR-1192, mmu-miR-377-3p, mmu-miR-27b-3p, and mmu-miR-146b-5p) due to low or undetectable expression in the kidney or unavoidable technical problems. Among the remaining 29 miRNAs, qPCR data demonstrated the differential expression of 13 and 6 miRNAs in the cortex and medulla, respectively (Fig. 3d).

A holistic miRNA-target interaction map was constructed, and key functions were inferred. To investigate the role of differentially expressed miRNAs, the validated targets were identified as encompassing 108 and 56 genes for cortex and medulla, respectively (Table S4). Then, the interactions between these genes and their targeting miRNAs were mapped (Fig. 4a, b). As expected for biological networks[77], both cortex and medulla networks followed a power law distribution. Graph theory measures such as degree, betweenness cen-

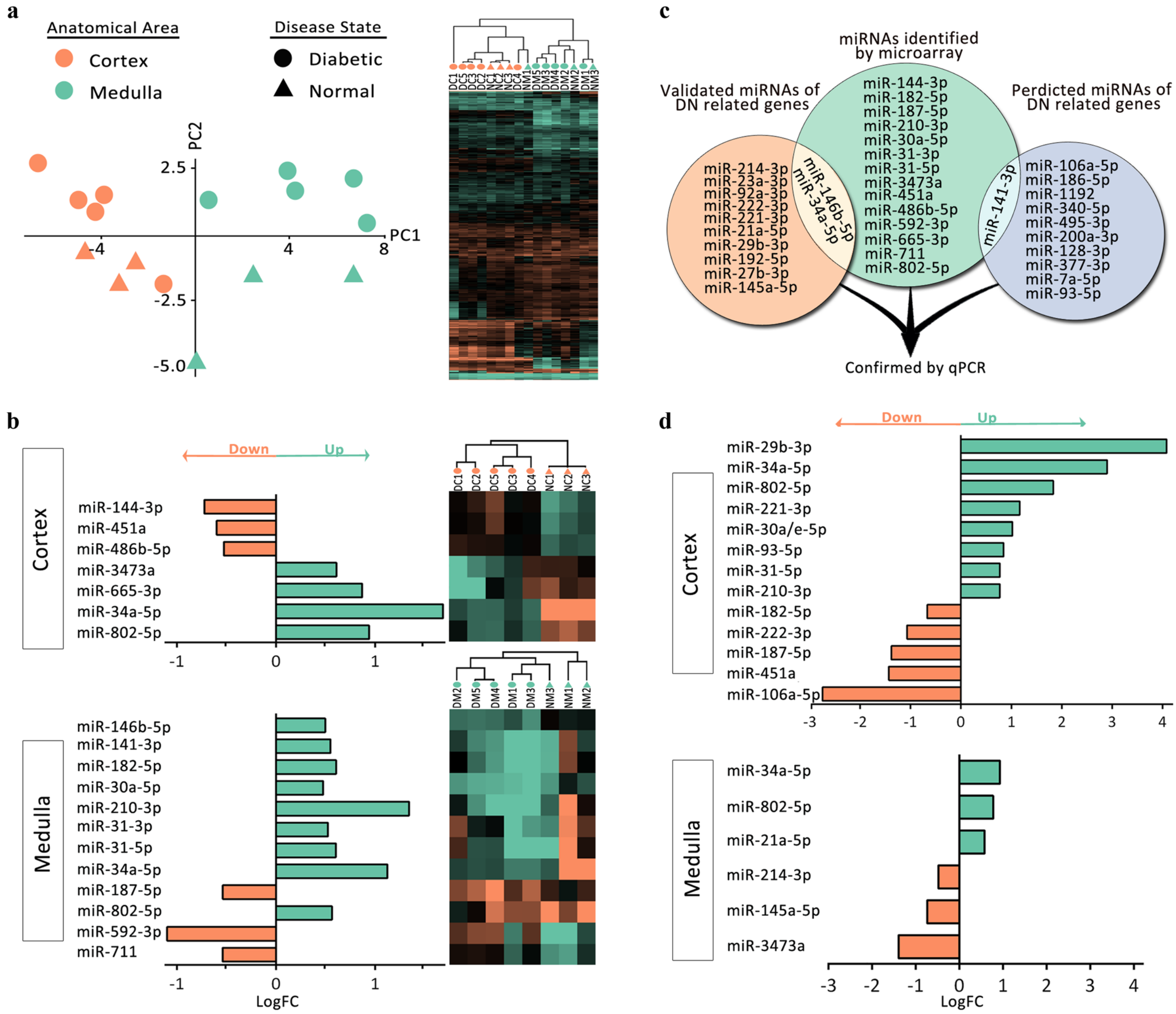

Figure 3. The kidney miRNA profile in DN. To assess microarray data quality in an unsupervised manner, principal component analysis and hierarchical clustering were performed (**a**). In microarray, miRNAs with |logFC|≥0.5 in cortex and medulla were determined (**b**). In addition to the miRNAs detected by microarray, some miRNAs which were experimentally shown or predicted to target DN-associated genes were selected (**c**). Among these candidates, 13 and 6 miRNAs were differentially expressed by qPCR in the cortex and medulla, respectively (**d**).

trality, and closeness centrality were determined to identify central nodes in the networks. The top genes in terms of these centrality parameters were assumed as central (Table S5). Interestingly, the critical role of the majority of the central nodes such as Hif1a, Vegfa, Sirt1, and Foxo1 has been shown in previous studies[78–81]. Similarly, there are experimental supports for the association of DN and miR-29, miR-34, miR-21, and miR-451, which we identified as central miRNAs[82–85]. This finding is in agreement with the concept that central network nodes drive critical functions[86].

To identify key interactions in the cortex and medulla networks, modules were determined as sub-graphs of dense interactions (Fig. 4c). Two modules were related to the extra-cellular matrix aligned with DN histopathology[87,88]. Notably, miR-29, a key player in DN[89,90] regulates all elements of these two modules. Another module is related to the epigenetic control of Sirt1 on Foxo3 and Stat3, which are shown to be associated with DN pathogenesis[91–93]. Pathway enrichment analysis was performed to identify the signaling pathways associated with the miRNA validated targets. Using KEGG and Reactome databases, forty and eleven inter-connected pathways were enriched for cortex (Fig. 5a) and medulla (Fig. 5b), respectively. Further pathways, as well as GO biological process, were also identified using BioCarta and GO consortium for cortex (Table S6) and medulla (Table S7). The role of most enriched pathways, including TGFB, FGFR, EGFR, Notch, and hypoxia signaling has been shown in DN[94–99]. This analysis was validated by pathway enrichment analysis for ten random gene lists of similar sizes which yielded no or a few unrelated signaling pathways (data not shown).

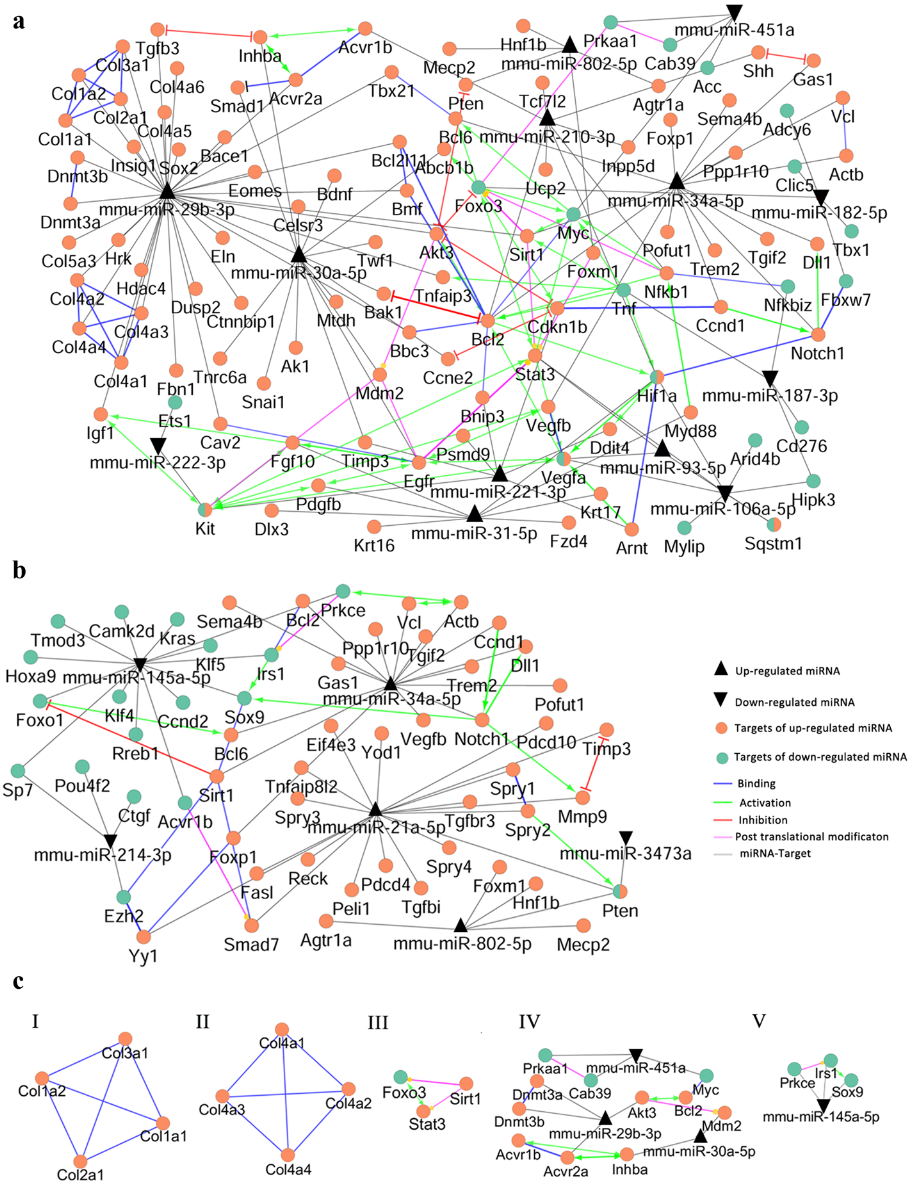

Figure 4. The holistic miRNA-target interaction maps were constructed, and key modules were identified. To investigate the role of differentially expressed miRNAs, the validated targets were identified, and the interaction networks were constructed with the differentially expressed miRNAs and their validated targets in the cortex (**a**) and medulla (**b**). Four modules in the cortex (I-IV) and one in medulla (V) networks were found, which potentially represent key interactions (**c**).

A novel high-performance machine learning method was developed to predict potential drug targets.

One of the main objectives of this research was to translate the findings on DN molecular pathogenesis into clinical applications. Thus, a machine learning approach was developed to predict which role player molecules could be suitable therapeutic targets. We followed the hypothesis that FDA-approved drug targets have unique properties compared to other proteins and these characteristics can be used to propose novel drug targets. Sixty-five biochemical features (Table S8) were determined which were harvested from several databases, for all human proteins (#20,132). Moreover, 23 network topology features (Table S9) were determined for the proteins in the human interactome obtained from HPRD (#9226). Based on DrugBank, 1443 proteins were then determined to be the targets of at least one FDA-approved drug. These proteins were considered as the DT group while the remaining 18,689 proteins were assigned to the non-DT class. Although the frequency or median of most biochemical (Fig. S2) and topology (Fig. S3) features are statistically different between DT and non-DT proteins, extensive distribution overlaps can be observed between these two protein classes. To assess the predictive value of these features for discrimination between DT and non-DT, different standard machine learning algorithms, namely logistic regression (LG), radial basis function kernel support vector machine (RBF-SVM), generalized linear model (GLM), and radial basis function network (RBFN) were utilized. The hold-out method

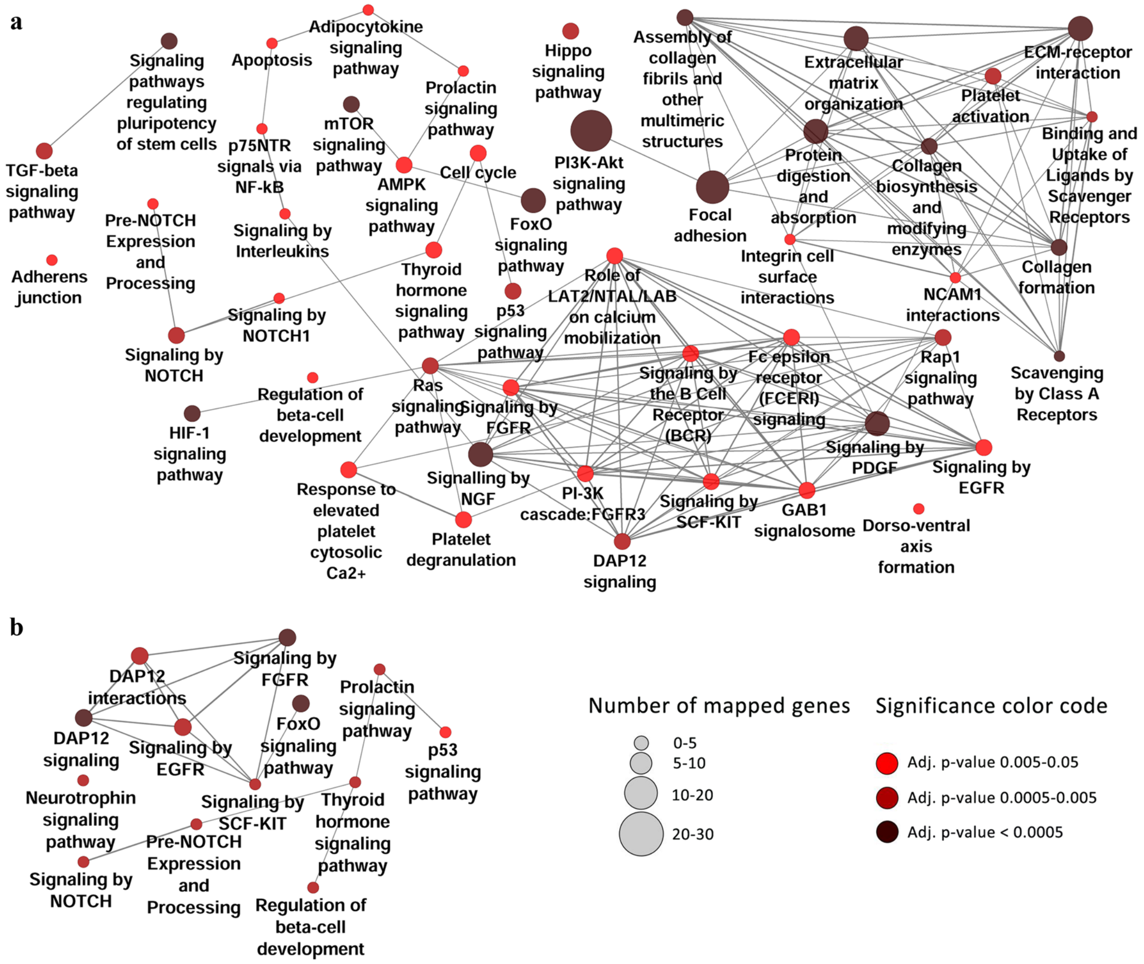
a
Adipocytokine signaling pathway
Apoptosis
Signaling pathways regulating pluripotency of stem cells
Hippo signaling pathway
Assembly of collagen fibrils and other multimeric structures
Extracellular matrix organization
ECM-receptor interaction
Platelet activation
Prolactin signaling pathway
mTOR signaling pathway
p75NTR signals via NF-kB
TGF-beta signaling pathway
PI3K-Akt signaling pathway
AMPK signaling pathway
Cell cycle
Protein digestion and absorption
Collagen biosynthesis and modifying enzymes
Binding and Uptake of Ligands by Scavenger Receptors
Focal adhesion
Pre-NOTCH Expression and Processing
Signaling by Interleukins
FoxO signaling pathway
Integrin cell surface interactions
Collagen formation
Adherens junction
Thyroid hormone signaling pathway
Role of LAT2/NTAL/LAB on calcium mobilization
NCAM1 interactions
p53 signaling pathway
Signaling by NOTCH1
Signaling by NOTCH
Fc epsilon receptor (FCERI) signaling
Rap1 signaling pathway
Scavenging by Class A Receptors
Signaling by the B Cell Receptor (BCR)
Ras signaling pathway
Signaling by FGFR
Regulation of beta-cell development
HIF-1 signaling pathway
Signaling by PDGF
Signaling by EGFR
Signalling by NGF
Response to elevated platelet cytosolic Ca2+
PI-3K cascade:FGFR3
Signaling by SCF-KIT
GAB1 signalosome
Dorso-ventral axis formation
Platelet degranulation
DAP12 signaling
b
DAP12 interactions
Signaling by FGFR
Prolactin signaling pathway
FoxO signaling pathway
DAP12 signaling
Signaling by EGFR
p53 signaling pathway
Neurotrophin signaling pathway
Signaling by SCF-KIT
Thyroid hormone signaling pathway
Pre-NOTCH Expression and Processing
Signaling by NOTCH
Regulation of beta-cell development
Number of mapped genes
0-5
5-10
10-20
20-30
Significance color code
Adj. p-value 0.005-0.05
Adj. p-value 0.0005-0.005
Adj. p-value < 0.0005

Figure 5. Key signaling pathways associated with DN. Pathway enrichment analysis with miRNA validated targets revealed 40 and 11 inter-connected pathways (adjusted $P \leq 0.05$) for cortex (**a**) and medulla (**b**), respectively.

measurements showed that the performance of none of the exploited methods was satisfying, neither with biochemical nor topology features (Fig. 6).

The failure of classical machines could be due to the unique nature of the data such as the considerable imbalanced size of the groups, significant overlap in the distribution of features in the two protein classes, and high dimensionality. The mentioned limitations were addressed by developing mGMDH-AFS, a high-performance machine learning method that considers high-level feature interactions. The performance of this novel tool was acceptable and significantly superior to the other standard machines, as revealed by hold-out validation (Fig. 6). The prediction power of mGMDH-AFS was further evaluated by fourfold cross-validation (Fig. S4). While topology features alone were not predictive for discrimination between DT and non-DT classes, biochemical features and the combination of biochemical and topology features were informative for the classification. The core topology features selected by mGMDH-AFS were included network degree, betweenness, and closeness. Also, being an enzyme, a receptor, an ion channel, or having post-translational modifications (PTM) were amongst the main biochemical features. To further assess the functionality of mGMDH-AFS, all human proteins were randomly labeled as 1 and 0 with the ratio of 1/0 the same as DT/non-DT classes. This procedure was repeated ten times to generate ten sets of randomly allocated proteins. As expected, the machine performance was significantly better with the real set compared to these random sets, indicating a reliable classification (Fig. S5).

After validating the mGMDH-AFS performance for all human proteins, this classifier was applied to predict potential DT in the constructed cortex and medulla networks. The algorithm calculated the probability of being DT for each protein in the networks. Using DrugBank to identify currently approved targets in these networks, an ROC curve analysis was carried out to evaluate machine performance (Fig. 7a). The top predicted drug targets for both cortex and medulla are Agtr1a, Egfr, Kit, Celsr3, Clic5, Tgfbr3, Acvr1b (Fig. 7b). Interestingly, the machine predicted Angiotensin II receptor, a well-known current drug target in DN, as the best candidate in the medulla and the 3rd best one in the cortex, supporting the validity of the developed approach. Some experimental supports were also found for other proposed targets such as EGF receptor and protein kinase C isoforms in previous studies on DN[100, 101].

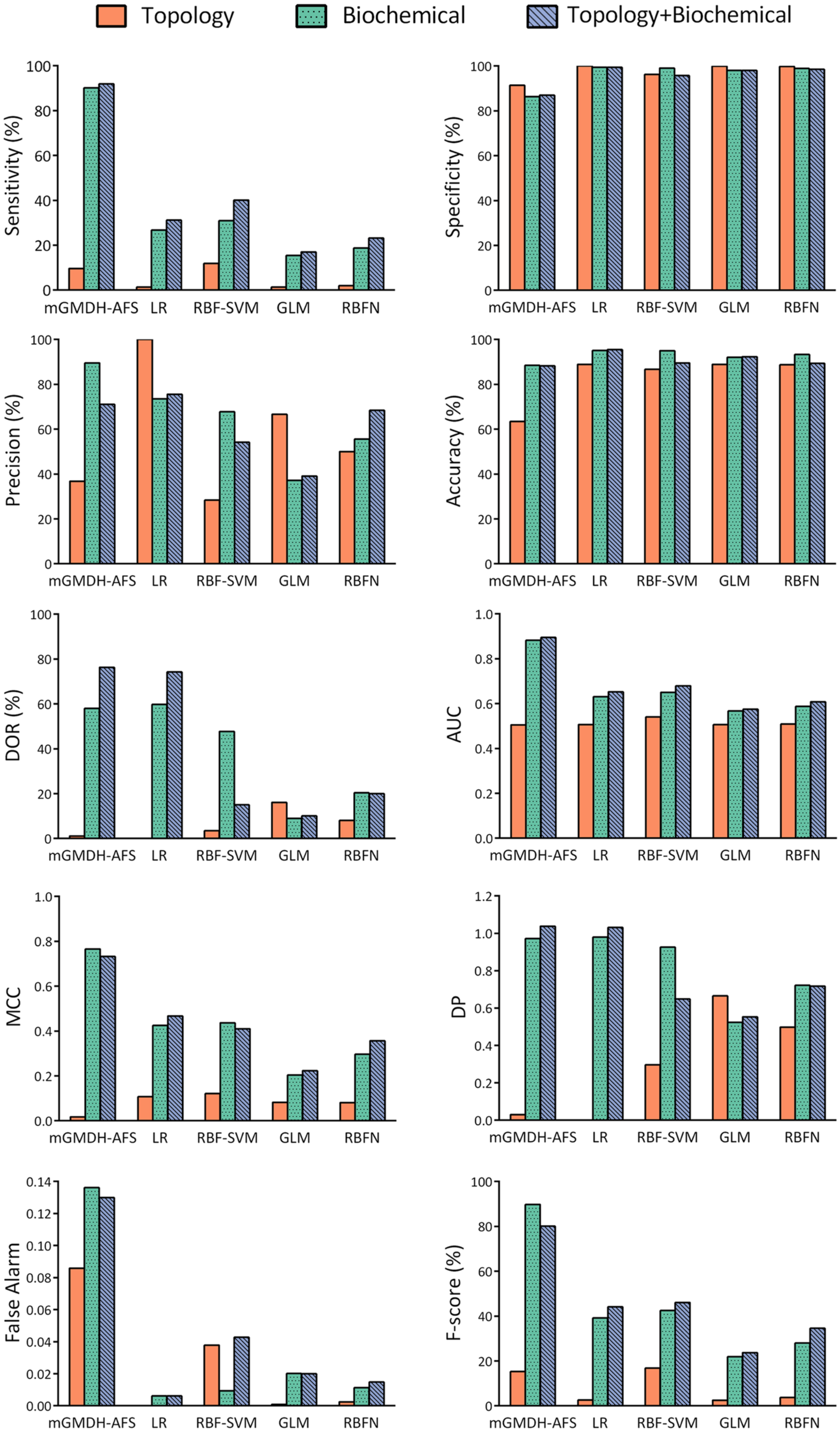

Figure 6. The novel machine could appropriately classify human proteins as drug targets or non-drug targets. The performance of the proposed mGMDH-AFS machine for drug target prediction based on biochemical or topology + biochemical features was acceptable and significantly superior to the examined standard machines including logistic regression (LG), radial basis function kernel support vector machine (RBF-SVM), generalized linear model (GLM), and radial basis function network (RBFN) as revealed by the hold-out validation. The proposed method significantly outperformed the state-of-the-art models (adjusted $P \leq 0.05$). AUC: area under the receiver operating characteristic (ROC) curve; DOR: diagnosis odds ratio; MCC: Matthews correlation coefficient; DP: discriminant power.

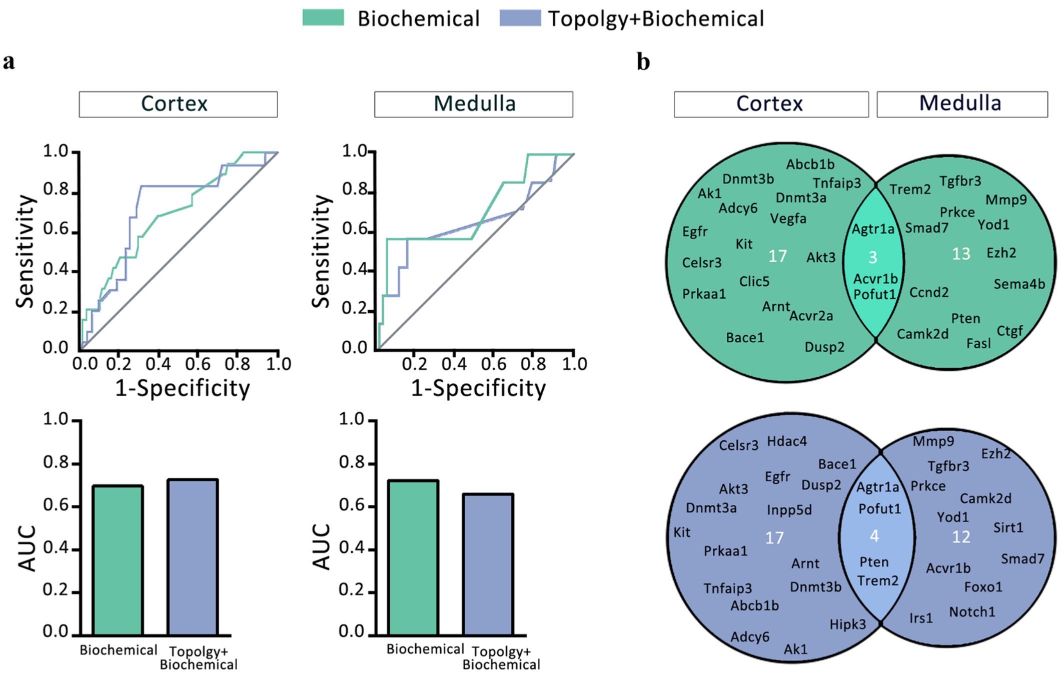

Figure 7. The developed machine learning algorithm proposed novel therapeutic targets for DN. The performance of the mGMDH-AFS classifier to predict drug targets in the constructed networks of cortex and medulla was assessed by ROC curve analysis (**a**). The top predicted targets in cortex and medulla networks are shown either with biochemical or topology + biochemical features (**b**).

Discussion

Despite considerable efforts, diabetic patients progress to end-stage renal disease at alarming proportions, necessitating further investigations to design more efficient therapeutic approaches[102]. Numerous studies have shown the role of individual elements in the pathogenesis of DN but systematic evaluations have rarely been attempted. In this study, miRNA-targets interaction maps were established in DN using system biology approaches. The central nodes, modules, and critical signaling pathways were determined. To translate the research findings to clinical applications, an innovative high-performance classifier was also developed to predict novel therapeutic targets in DN.

The selection of miRNAs as the starting point for investigating DN molecular pathogenesis was based on the concept that they target functionally related genes. Indeed, each variably expressed miRNA can be a clue to identify a group of related altered functions[5,76]. Microarray profiling and bioinformatics predictions were employed in parallel to identify some miRNAs for further validation with qPCR. Among the identified DE miRNAs, the roles of miR-34a, miR-29b, miR-21a, miR145a, and miR-451a have been extensively shown in DN[83,84,89,103,104], indicating the validity of our approach. However, to the best of our knowledge, miR-802-5p, 182-5p, 210-3p, 31-5p, and 106a-5p are for the first time attributed to DN in the current study.

One of the advantages of the present study is separate miRNA profiling for the cortex and medulla. Remarkably, the PCA plot demonstrated that the effect of anatomical area on miRNA expression was higher than the impact of the disease state. Although miR-802-5p and miR-34a are overexpressed in both cortex and medulla of DN mice, the other identified miRNAs are differential in either cortex or medulla. This finding is in line with recent studies revealing diverse mRNA and miRNA expression profiles for different anatomical parts of the normal kidney[105,106].

To provide evidence about the role of the identified miRNAs, their targets were collected and miRNA-target interaction networks were constructed. The constructed networks allowed a global systematic view towards the collaborations between miRNAs and their interactions with corresponding targets. Considering the complexity of biological interaction networks, it is critical to identify the elements with the highest influence on the outcome. Therefore, we performed topology analysis to determine the central nodes in the networks, as it has been previously shown that they drive key signaling pathways[86]. Interestingly, the in vivo knockdown of miR-29, one of the central miRNAs in the cortex network, has been shown to halt the progression of DN[82]. Notably, in the cortex network, this miRNA regulates two modules composed of extracellular matrix (ECM) elements. The identification of two modules associated with ECM in the cortex network is aligned with the histopathologic finding of fibrosis in the cortex.

The role of DE miRNAs was further explored by performing pathway enrichment analysis to determine the signaling pathways associated with miRNA targets. Most of the enriched pathways, including TGFB, EGFR, and Notch signaling pathways were experimentally shown to be associated with DN[107–109]. Using this analysis, we could also predict novel pathways whose roles in DN remain to be confirmed in future studies. NGF signaling pathway which is related to diabetic neuropathy was amongst the enriched pathways[110]. Similarly, the platelet degranulation pathway was detected as a potential role player in DN. Previous studies have demonstrated the role of this pathway in some pro-fibrotic disorders such as idiopathic pulmonary fibrosis and myelofibrosis[111,112].

The findings of this research were translated to clinical applications. It is hypothesized that the current FDA-approved drugs affect proteins with distinctive properties and these features can potentially be used to introduce novel drugs. Although some studies have shown that current DTs have unique features[8,113], the construction of classifiers has remained a major challenge due to the overlap of feature distributions. To distinguish these two protein classes, some previous investigators have employed common machine learning approaches including LR, SVM, GLM, and RBFN[9,14,48,50]. In this study, however, these methods did not lead to satisfying results. This discrepancy could be assigned to the fact that these studies used equal size DT and non-DT classes to increase machine performance, or assessed the outputs with limited indices.

To address the limitations of current classifiers, a high-performance machine learning method was developed. Unlike other methods, we considered the original highly imbalanced datasets for classification. However, learning from the imbalanced data is challenging[114]. In this algorithm, a cost function other than the traditional mean-square-error metrics was utilized to avoid learning bias toward the majority class. The fitness function of the Matthews correlation coefficient was used whose robustness in imbalanced datasets has been proven[71,115].

In addition to the initial feature selection using the Relief algorithm, different interactions of the selected input features were considered during GMDH deep learning procedure. Indeed, biological systems are very complex and the nonlinear interactions of the non-redundant features could improve the performance of their classification systems[116]. Based on different assessments, the performance of mGMDH-AFS is superior to that of the state-of-the-art approaches, suggesting it as a promising approach for therapeutic target prediction in complex disorders.

After several validation steps, the developed machine learning algorithm was employed to analyze the cortex and medulla networks. It provided well-balanced diagnostic accuracy rates and resulted in a list of novel promising therapeutic targets for DN, which can be assessed in upcoming investigations. Among the top-ranked candidates in the cortex or medulla are Agtr1a, Egfr, Clic5, and Prkce. Interestingly, Agtr1a is the target of angiotensin receptor blockers which are currently in the market for DN treatment[117]. It has been also shown that the inhibition of Egfr or Prkc isoforms by small molecules can prevent the progression of DN[100,101].

In conclusion, a combination of experimental and computational methods was exploited to generate a holistic map of DN and introduce novel therapeutic targets. The limitation of this work is the restriction of experimental data to miRNA profiling. As a future perspective, we plan to integrate other omics layers into the constructed networks to achieve more accurate insights. Notably, the proposed approach for drug target prediction could also be employed for other complex disorders as well.

Received: 3 March 2021; Accepted: 12 November 2021
Published online: 06 December 2021

References

1. Gembillo, G. *et al.* Kidney disease in diabetic patients: from pathophysiology to pharmacological aspects with a focus on therapeutic inertia. *Int. J. Mol. Sci.* **22**(9), 4824 (2021).
2. Brosius, F. C. & Ju, W. The promise of systems biology for diabetic kidney disease. *Adv. Chronic Kidney Dis.* **25**(2), 202–213 (2018).
3. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Brief Bioinform.* **19**(6), 1370–1381 (2018).
4. Pranavkrishna, S., Sanjeev, G., Akshaya, R. L., Rohini, M. & Selvamurugan, N. A computational approach on studying the regulation of TGF-β1-stimulated Runx2 expression by MicroRNAs in human breast cancer cells. *Comput. Biol. Med.* **137**, 104823 (2021).
5. Tseng, C. W., Lin, C. C., Chen, C. N., Huang, H. C. & Juan, H. F. Integrative network analysis reveals active microRNAs and their functions in gastric cancer. *BMC Syst Biol.* **5**, 99 (2011).
6. Tang, J. *et al.* The role of MicroRNAs in the pathogenesis of diabetic nephropathy. *Int. J. Endocrinol.* **2019**, 8719060 (2019).
7. Kiyanpour, F., Abedi, M. & Gheisari, Y. A systematic integrative approach reveals novel microRNAs in diabetic nephropathy. *J. Res. Med. Sci.* **25**, 1 (2020).
8. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1**(9), 727–730 (2002).
9. Jeon, J. *et al.* A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome medicine.* **6**(7), 57 (2014).
10. Kumari, P., Nath, A. & Chaube, R. Identification of human drug targets using machine-learning algorithms. *Comput. Biol. Med.* **56**, 175–181 (2015).
11. Li, Z. C. *et al.* Large-scale identification of potential drug targets based on the topological features of human protein-protein interaction network. *Anal. Chim. Acta* **871**, 18–27 (2015).
12. Bakheet, T. M. & Doig, A. J. Properties and identification of human protein drug targets. *Bioinformatics* **25**(4), 451–457 (2009).
13. Bull, S. C. & Doig, A. J. Properties of protein drug target classes. *PLoS ONE* **10**(3), e0117955 (2015).
14. Yao, L. & Rzhetsky, A. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res.* **18**(2), 206–213 (2008).
15. Zhu, M. *et al.* The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J. Drug Target.* **17**(7), 524–532 (2009).
16. Guide for the Care and Use of Laboratory Animals, 8th edn, 2011.
17. Morris, J. H. *et al.* clusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinform.* **12**, 436 (2011).
18. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003).
19. Wickham, H. *ggplot2 Elegant Graphics for Data Analysis* (Springer, New York, 2009).
20. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014).
21. Pfaffl, M. W., Horgan, G. W. & Dempfle, L. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res.* **30**(9), e36 (2002).

22. Chou, C. H. *et al.* miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **46**(D1), D296–D302 (2018).
23. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, 05005 (2015).
24. Dweep, H. & Gretz, N. miRWalk2.0: A comprehensive atlas of microRNA-target interactions. *Nat. Methods* **12**(8), 697 (2015).
25. Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. *Bioinformatics* **29**(5), 661–663 (2013).
26. Szklarczyk, D. *et al.* The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**(D1), D362–D368 (2017).
27. Assenov, Y., Ramirez, F., Schelhorn, S. E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**(2), 282–284 (2008).
28. Tang, Y., Li, M., Wang, J., Pan, Y. & Wu, F. X. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems* **127**, 67–72 (2015).
29. Scardoni, G. *et al.* Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Res* **3**, 139 (2014).
30. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **4**(1), 2 (2003).
31. Keshava Prasad, T. S. *et al.* Human protein reference database–2009 update. *Nucleic Acids Res.* **37**(Database issue), D767–D772 (2009).
32. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**(8), 1091–1093 (2009).
33. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**(D1), D649–D655 (2018).
34. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**(D1), D545–D551 (2021).
35. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* **4**(1), 44–57 (2009).
36. Law, V. *et al.* DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **42**(Database issue), D1091–D1097 (2014).
37. Apweiler, R. B. A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**(D1), D158–D169 (2017).
38. Huang, K. Y. *et al.* dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.* **44**(D1), D435–D446 (2016).
39. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**(1), 304–305 (2000).
40. Corcoran, C. C., Grady, C. R., Pisitkun, T., Parulekar, J. & Knepper, M. A. From 20th century metabolic wall charts to 21st century systems biology: Database of mammalian metabolic enzymes. *Am. J. Physiol. Renal Physiol.* **312**(3), F533–F542 (2017).
41. Chawla, K., Tripathi, S., Thommesen, L., Laegreid, A. & Kuiper, M. TFcheckpoint: A curated compendium of specific DNA-binding RNA polymerase II transcription factors. *Bioinformatics* **29**(19), 2519–2520 (2013).
42. Zhang, H. M. *et al.* AnimalTFDB: A comprehensive animal transcription factor database. *Nucleic Acids Res.* **40**(100), D144–D149 (2012).
43. Medvedeva, Y. A. *et al.* EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015**, bav067 (2015).
44. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
45. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**(D1), D1251–D1257 (2016).
46. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**(Database issue), D789–D898 (2015).
47. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**(D1), D896–D901 (2017).
48. Costa, P. R., Acencio, M. L. & Lemke, N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics* **11**(Suppl 5), S9 (2010).
49. Emig, D. *et al.* Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE* **8**(4), e60618 (2013).
50. Mitsopoulos, C., Schierz, A. C., Workman, P. & Al-Lazikani, B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLoS Comput. Biol.* **11**(12), e1004597 (2015).
51. Webb, A. R. & Copsey, K. D. *Introduction to Statistical Pattern Recognition* 3rd edn. (Wiley, New York, 2011).
52. Freedman, D. A. *Statistical Models: Theory and Practice* 2nd edn. (Cambridge University Press, 2009).
53. Wu, K. P. & Wang, S. D. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. *Pattern Recognit.* **42**(5), 710–717 (2009).
54. Peter Bartlett, K. P. B. *et al.* Advances in Kernel Methods. In *Support Vector Learning* (eds BernhardSchölkopf, C. J. C. B. & Smola, A. J.) (The MIT Press, 1998).
55. Collett, D. *Modelling Binary Data* (Chapman and Hall/CRC, 2002).
56. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* 2nd edn. (Chapman and Hall/CRC, 1989).
57. Schwenker, F., Kestler, H. A. & Palm, G. Three learning phases for radial-basis-function networks. *Neural Netw.* **14**(4–5), 439–458 (2001).
58. Abdel-Aal, R. E. GMDH-based feature ranking and selection for improved classification of medical data. *J. Biomed. Inform.* **38**(6), 456–468 (2005).
59. Marateb, H. R. *et al.* Prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in children and adolescents: The CASPIAN-III study. *Comput. Struct. Biotechnol. J.* **16**, 121–130 (2018).
60. Bennett MOMJEFACMBLLKP. Feature selection for in-silico drug design using genetic algorithms and neural networks. Soft Computing in Industrial Applications, 2001 SMCia/01 Proceedings of the 2001 IEEE Mountain Workshop on. (2001).
61. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* **11**(9), 647–657 (2010).
62. Clerc, M. *Particle Swarm Optimization* (Wiley, 2010).
63. Deng, Z., Chung, F.-L. & Wang, S. Robust relief-feature weighting, margin maximization, and fuzzy optimization. *IEEE Trans. Fuzzy Syst.* **18**(4), 726–744 (2010).
64. Bender, R. & Grouven, U. Using binary logistic regression models for ordinal data with non-proportional odds. *J. Clin. Epidemiol.* **51**(10), 809–816 (1998).
65. Hedeker, D. MIXNO: A computer program for mixed-effects nominal logistic regression. J. Stat. Softw. (1999).
66. Green, P. J. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives. *J. R. Stat. Soc. Ser. B (Methodol.).* **46**(2), 149–170 (1984).

67. Sun, Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Trans Pattern Anal Mach Intell.* **29**(6), 1035–1051 (2007).
68. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
69. D'Addabbo, A. & Maglietta, R. Parallel selective sampling method for imbalanced and large data classification. *Pattern Recognit. Lett.* **62**, 61–67 (2015).
70. Ivakhnenko, A. G. Heuristic self-organization in problems of engineering cybernetics. *Automatica* **6**(2), 207–219 (1970).
71. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **12**(6), e0177678 (2017).
72. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **4**(2), 627–635 (2013).
73. Sammut, C. & Webb, G. I. *Encyclopedia of Machine Learning* (Springer US, 2011).
74. Mosteller, F. A k-sample slippage test for an extreme population. *Ann. Math. Stat.* **19**(1), 58–65 (1948).
75. Good, P. (ed.) *Permutation Test A Practical Guide to Resampling Methods for Testing Hypotheses* (Springer, New York, 1994).
76. Krishnan, K. *et al.* MicroRNA-182-5p targets a network of genes involved in DNA repair. *RNA* **19**(2), 230–242 (2013).
77. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**(2), 101–113 (2004).
78. Nayak, B. K. *et al.* HIF-1 mediates renal fibrosis in OVE26 type 1 diabetic mice. *Diabetes* **65**(5), 1387–1397 (2016).
79. Sivaskandarajah, G. A. *et al.* Vegfa protects the glomerular microvasculature in diabetes. *Diabetes* **61**(11), 2958–2966 (2012).
80. Hasegawa, K. *et al.* Renal tubular Sirt1 attenuates diabetic albuminuria by epigenetically suppressing Claudin-1 overexpression in podocytes. *Nat Med.* **19**(11), 1496–1504 (2013).
81. Li, W. *et al.* FoxO1 promotes mitophagy in the podocytes of diabetic male mice via the PINK1/Parkin pathway. *Endocrinology* **158**(7), 2155–2167 (2017).
82. Long, J., Wang, Y., Wang, W., Chang, B. H. & Danesh, F. R. MicroRNA-29c is a signature microRNA under high glucose conditions that targets Sprouty homolog 1, and its in vivo knockdown prevents progression of diabetic nephropathy. *J. Biol. Chem.* **286**(13), 11837–11848 (2011).
83. Zhang, L. *et al.* Down-regulation of miR-34a alleviates mesangial proliferation in vitro and glomerular hypertrophy in early diabetic nephropathy mice by targeting GAS1. *J. Diabetes Comp.* **28**(3), 259–264 (2014).
84. Zhong, X. *et al.* miR-21 is a key therapeutic target for renal injury in a mouse model of type 2 diabetes. *Diabetologia* **56**(3), 663–674 (2013).
85. Sun, Y. *et al.* miR-451 suppresses the NF-kappaB-mediated proinflammatory molecules expression through inhibiting LMP7 in diabetic nephropathy. *Mol. Cell Endocrinol.* **433**, 75–86 (2016).
86. Abedi, M. & Gheisari, Y. Nodes with high centrality in protein interaction networks are responsible for driving signaling pathways in diabetic nephropathy. *PeerJ* **3**, e1284 (2015).
87. Tervaert, T. W. *et al.* Pathologic classification of diabetic nephropathy. *J. Am. Soc. Nephrol.* **21**(4), 556–563 (2010).
88. Mason, R. M. & Wahab, N. A. Extracellular matrix metabolism in diabetic nephropathy. *J. Am. Soc. Nephrol.* **14**(5), 1358–1373 (2003).
89. Chen, H. Y. *et al.* MicroRNA-29b inhibits diabetic nephropathy in db/db mice. *Mol. Ther.* **22**(4), 842–853 (2014).
90. Wang, B. *et al.* Suppression of microRNA-29 expression by TGF-beta1 promotes collagen expression and renal fibrosis. *J. Am. Soc. Nephrol.* **23**(2), 252–265 (2012).
91. Kong, L. *et al.* Sirtuin 1: A target for kidney diseases. *Mol. Med.* **21**, 87–97 (2015).
92. Ding, Y. & Choi, M. E. Autophagy in diabetic nephropathy. *J. Endocrinol.* **224**(1), R15-30 (2015).
93. Nie, Y. *et al.* STAT3 inhibition of gluconeogenesis is downregulated by SirT1. *Nat. Cell Biol.* **11**(4), 492–500 (2009).
94. Koya, D. Epidermal growth factor receptor signaling and the progression of diabetic nephropathy. *J. Diabetes Investig.* **6**, 519–521 (2015).
95. Bonegio, R. & Susztak, K. Notch signaling in diabetic nephropathy. *Exp. Cell Res.* **318**(9), 986–992 (2012).
96. Yamamoto, T., Nakamura, T., Noble, N. A., Ruoslahti, E. & Border, W. A. Expression of transforming growth factor beta is elevated in human and experimental diabetic nephropathy. *Proc. Natl. Acad. Sci.* **90**(5), 1814–1818 (1993).
97. Tsuchida, K. *et al.* Suppression of transforming growth factor beta and vascular endothelial growth factor in diabetic nephropathy in rats by a novel advanced glycation end product inhibitor, OPB-9195. *Diabetologia* **42**(5), 579–588 (1999).
98. Nordquist, L. *et al.* Activation of hypoxia-inducible factors prevents diabetic nephropathy. *J. Am. Soc. Nephrol.* **26**(2), 328–338 (2015).
99. Wyatt, C. M. & Drueke, T. B. Fibroblast growth factor receptor 4: The missing link between chronic kidney disease and FGF23-induced left ventricular hypertrophy?. *Kidney Int.* **89**(1), 7–9 (2016).
100. Xu, Z. *et al.* EGFR inhibition attenuates diabetic nephropathy through decreasing ROS and endoplasmic reticulum stress. *Oncotarget* **8**(20), 32655–32667 (2017).
101. Menne, J. *et al.* Dual inhibition of classical protein kinase C-alpha and protein kinase C-beta isoforms protects against experimental murine diabetic nephropathy. *Diabetes* **62**(4), 1167–1174 (2013).
102. Eboh, C. & Chowdhury, T. A. Management of diabetic renal disease. *Ann. Transl. Med.* **3**(11), 154 (2015).
103. Barutta, F. *et al.* Urinary exosomal microRNAs in incipient diabetic nephropathy. *PLoS ONE* **8**, e73798 (2013).
104. Zhang, Z. *et al.* MicroRNA-451 regulates p38 MAPK signaling by targeting of Ywhaz and suppresses the mesangial hypertrophy in early diabetic nephropathy. *FEBS Lett.* **586**(1), 20–26 (2012).
105. Higgins, J. P. *et al.* Gene expression in the normal adult human kidney assessed by complementary DNA microarray. *Mol. Biol. Cell* **15**(2), 649–656 (2004).
106. Tian, Z., Greene, A. S., Pietrusz, J. L., Matus, I. R. & Liang, M. MicroRNA–target pairs in the rat kidney identified by microRNA microarray, proteomic, and bioinformatic analysis. *Genome Res.* **18**(3), 404–411 (2008).
107. Yang, C. *et al.* SMAD3 promotes autophagy dysregulation by triggering lysosome depletion in tubular epithelial cells in diabetic nephropathy. *Autophagy* **17**, 2325–2344 (2020).
108. Sheng, L., Bayliss, G. & Zhuang, S. Epidermal growth factor receptor: A potential therapeutic target for diabetic kidney disease. *Front. Pharmacol.* **11**, 598910 (2020).
109. Wang, Y., Chang, J., Wang, Z. Q. & Li, Y. Sirt3 promotes the autophagy of HK-2 human proximal tubular epithelial cells via the inhibition of Notch-1/Hes-1 signaling. *Mol. Med. Rep.* **24**(3), 12273 (2021).
110. Li, K. *et al.* Taurine protects against myelin damage of sciatic nerve in diabetic peripheral neuropathy rats by controlling apoptosis of schwann cells via NGF/Akt/GSK3β pathway. *Exp Cell Res.* **383**(2), 111557 (2019).
111. Carrington, R., Jordan, S., Wong, Y. J., Pitchford, S. C. & Page, C. P. A novel murine model of pulmonary fibrosis: the role of platelets in chronic changes induced by bleomycin. *J. Pharmacol. Toxicol. Methods.* **109**, 107057 (2021).
112. Melo-Cardenas, J., Migliaccio, A. R. & Crispino, J. D. The role of megakaryocytes in myelofibrosis. *Hematol. Oncol. Clin. North Am.* **35**(2), 191–203 (2021).
113. Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**(10), 1119–1126 (2007).

114. Sun, Y., Wong, A. K. C. & Kamel, M. S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **23**(04), 687–719 (2009).
115. Saghapour, E., Kermani, S. & Sehhati, M. A novel feature ranking method for prediction of cancer stages using proteomics data. *PLoS ONE* **12**(9), e0184203 (2017).
116. Wang, H., Lo, S. H., Zheng, T. & Hu, I. Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics* **28**(21), 2834–2842 (2012).
117. Ruggenenti, P. *et al.* Preventing microalbuminuria with benazepril, valsartan, and benazepril-valsartan combination therapy in diabetic patients with high-normal albuminuria: A prospective, randomized, open-label, blinded endpoint (PROBE) study. *PLoS Med.* **18**(7), e1003691 (2021).

Author contributions

M.A., S.H.A., and S.M.N. performed the experiments and acquired the data. H.R.M. designed and managed machine learning studies. M.R.M. performed machine learning methods. M.A. wrote the draft of the manuscript which was critically revised by the other authors. Y.G. designed and supervised the whole project. All authors contributed to data analysis, approved the final manuscript, and agreed to be accountable for all aspects of the work.

Funding

This study was supported by National Institute for Medical Research Development (NIMAD; grant no. 942907), and Isfahan University of Medical Sciences (grant no. 199503).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-02282-3.

Correspondence and requests for materials should be addressed to Y.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.