ORIGINAL ARTICLE

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

MAYO CLINIC

Acoustic Analysis and Prediction of Type 2 Diabetes Mellitus Using Smartphone-Recorded Voice Segments

Jaycee M. Kaufman, MSc; Anirudh Thommandram, MASc; and Yan Fossat, MSc

Abstract

Objective: To investigate the potential of voice analysis as a prescreening or monitoring tool for type 2 diabetes mellitus (T2DM) by examining the differences in voice recordings between nondiabetic and T2DM individuals.

Patients and Methods: Total 267 participants diagnosed as nondiabetic (79 women and 113 men) or T2DM (18 women and 57 men) on the basis of American Diabetes Association guidelines were recruited in India between August 30, 2021 and June 30, 2022. Using a smartphone application, participants recorded a fixed phrase up to 6 times daily for 2 weeks, resulting in 18,465 recordings. Fourteen acoustic features were extracted from each recording to analyze differences between nondiabetic and T2DM individuals and create a prediction methodology for T2DM status.

Results: Significant differences were found between voice recordings of nondiabetic and T2DM men and women, both in the entire dataset and in an age-matched and body mass index (BMI [calculated as the weight in kilograms divided by the height in meters squared])-matched sample. The highest predictive accuracy was achieved by pitch ($P<.0001$), pitch SD ($P<.0001$), and relative average pertubation jitter ($P=.02$) for women, and intensity ($P<.0001$) and 11-point amplitude perturbation quotient shimmer (apq11, $P<0.0001$) for men. Incorporating these features with age and BMI, the optimal prediction models achieved accuracies of 0.75 ± 0.22 for women and 0.70 ± 0.10 for men through 5-fold cross-validation in the age-matched and BMI-matched sample.

Conclusion: Overall, vocal changes occur in individuals with T2DM compared with those without T2DM. Voice analysis shows potential as a prescreening or monitoring tool for T2DM, particularly when combined with other risk factors associated with the condition.

Trial Registration: clinicaltrials.gov Identifier: CTRI/2021/08/035957

 ■ Mayo Clin Proc Digital Health 2023;1(4):534-544

From Klick Applied Sciences, Klick Inc, Toronto, Canada (J.M.K., A.T., Y.F.); and Faculty of Science, Ontario Tech University, Oshawa, Canada (Y.F.).

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by impaired insulin action and elevated blood glucose levels. Its rising prevalence and considerable effect on global health have gained substantial attention in recent years, prompting a push for proactive measures. An estimated 175 million individuals worldwide have undiagnosed diabetes, and the cumulative economic burden is estimated to reach nearly $2.1 trillion per year in 2030.[1] In addition, diabetes diagnosis is associated with an increased risk of mortality from cancer, renal disease, infections, liver disease, nervous system disorders, and chronic obstructive pulmonary disease.[2] It is imperative to develop effective strategies for disease detection that can identify individuals earlier in the disease trajectory, allowing for timely interventions and alleviating the consequences for individuals and health care infrastructures.

Recently, voice has emerged as a promising candidate for pathology detection and screening. It is noninvasive, inexpensive, and convenient, as voices can be recorded using a smartphone or portable device. Not only does this increase patient comfort, but it also has the potential for use in remote and underserved communities that may have limited access to health care services.

Voice synthesis is a complex process that relies on the combined effects of the

respiratory system, the nervous system, and the larynx. Anything that affects these systems can influence the voice, whether it is perceptible audibly or detectable through computer analysis.[3] In T2DM, individuals experience sustained periods of high blood glucose. Point-in-time glucose concentrations have been hypothesized to affect the elastic properties of the vocal chords,[4] and long-term elevated glucose can have detrimental effects such as peripheral neuropathy and myopathy (ie, the damage of nerve and muscle fibers, respectively).[5,6] Myopathy has been shown to correlate with an increased prevalence of voice disorders and dysphagia,[7] potentially because of muscle weakness within the larynx, whereas hoarseness, vocal straining, and aphonia are present in individuals with diabetic neuropathy.[8] Furthermore, T2DM has been linked to an increased prevalence of psychological disorders such as depression, anxiety, eating disorders, and decreasing cognitive function, all of which have been linked to vocal changes.[6,9-13] Overall, there is strong justification for vocal differences occurring in T2DM, and previous work has shown there are distinct vocal differences between T2DM and nondiabetic individuals.[14-16]

The objective of this manuscript is to assess the feasibility of voice for the prediction and detection of T2DM. Although there have been some promising preliminary results, there is limited data on vocal changes between nondiabetic and T2DM individuals in age-matched and basic metabolic index–matched populations, and analysis has yet to be performed on a fixed sentence despite the reported success in determining glucose-related voice changes from spoken sentences and free speech.[4] This manuscript aims to address the following points to identify vocal changes in T2DM:

1. Simulate real-world scenarios by collecting data using a mobile application and using a fixed phrase recording.
2. Perform analysis of voice features separately for men and women.
3. Perform analysis of voice features in age-matched and BMI-matched sample.
4. Use relevant voice features to predict T2DM status.
5. Combine existing risk factors (age and BMI) with voice results to predict T2DM status.

We first performed statistical analysis of the acoustic voice features to identify differences in the voices of T2DM men and women compared with healthy controls, in both the entire dataset and in an age-matched and BMI-matched dataset. We then used the extracted features and simple machine learning models to create a prediction methodology for the detection of T2DM. Finally, we incorporated age and BMI into the prediction results, increasing the accuracy of the proposed method. Overall, the success of the prediction model in this pilot study justifies the use of voice analysis in T2DM screening.

Previous Work

Vocal biomarker prediction of disease has been used for pathologies ranging from coronary artery disease to pulmonary function to Parkinson disease to COVID-19 detection.[17-22] Studies that have employed machine learning methods in voice pathology prediction range from using a large set of features (>6000 features) in the time and frequency domains,[19] to small feature sets (10 features).[18] Common prediction models include Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree, and Random Forest models.[18-20,22,23]

Looking specifically at diabetes, a recent study assessing vocal changes in cystic fibrosis–related diabetes analyzed 7 features: fundamental frequency, fundamental frequency variation, jitter, shimmer, noise-to-harmonic ratio, voice turbulence index, and relative average perturbation (RAP). This study used LR to determine the predictive capabilities of significant ($P<.05$) features.[24] In T2DM, previous studies have focused primarily on identifying features that may be different between the T2DM and nondiabetic populations, with varying results. A study of 83 participants conducted in Thailand indicated that fundamental frequency decreases significantly in T2DM women when compared with nondiabetic women.[14] However, this study did not

TABLE 1. Demographic Characteristic Information for Study Participants[a,b]

Characteristic	Entire Dataset			Matched Dataset		
	ND	T2DM	*P*	ND	T2DM	*P*
Woman						
Number of participants	79	18	—	11	11	—
Total number of recordings	5636	1183	—	503	505	—
Age (y)	32.66±10.85	28.20±9.52	<.001	45.73±10.47	45.91±10.85	.97
BMI, kg/m^2	28.20±9.52	35.09±11.36	.01	29.09±5.29	31.41±5.40	.32
Number of recordings per participant	71.34±23.75	65.72±25.39	.37	77.54±10.56	70.91±20.39	.35
Man						
Number of participants	113	57	—	29	29	—
Total number of recordings	7935	3711	—	1964	2047	—
Age (y)	32.73±11.36	48.96±10.03	<.001	45.00±10.92	44.76±10.61	0.38
BMI, kg/m^2	27.69±7.54	30.87±14.33	<.001	27.77±4.46	26.70±3.28	0.19
Number of recordings per participant	70.22±23.23	65.11±36.78	0.27	67.72±33.27	70.59±25.40	.72

[a]BMI, body mass index; ND, nondiabetic; T2DM, type 2 diabetic mellitus.
[b]Values are displayed as mean ± SD. *P*-values from independent 2-tailed *t* test. Statistical significance ($P<.05$).

identify any differences in men, a result confirmed by a previous study conducted in 2012.[25] On the contrary, a study conducted in 2021 on 51 people with diabetes indicated that individuals with T2DM had an increased absolute jitter value compared with their healthy controls, although the sample was not segmented into male and female categories for the analysis.[15] Voice analysis conducted on 177 voice samples in 2016 found a decrease in all vocal parameters for women with T2DM and all vocal parameters for men with T2DM except absolute jitter and RAP.[16] All previous studies in T2DM analyzed sustained phonation of the vowel a. Features used in analysis were fundamental frequency (3 studies), harmonic noise ratio (4 studies), jitter (3 studies), shimmer (4 studies), RAP jitter (3 studies), amplitude perturbation quotient (APQ) shimmer (2 studies), phonation time (2 studies), and voice turbulence index (1 study).

METHODS

Participants and Study Design

Participants were recruited as part of a larger study involving the relationship between voice and glucose control between August 30, 2021 and June 30, 2022. Participants were recruited from 4 sites in India and diagnosed by a physician as nondiabetic or type 2 diabetic according to guidelines set by the American Diabetes Association (ADA).[26] All participants signed informed consent. The study received full ethics clearance, and all methods were conducted in accordance with relevant guidelines and regulations. Participants were instructed to record their voice at least 6 times a day for 2 weeks into a custom mobile application, saying the fixed sentences "Hello, how are you? What is my glucose level right now?" Voice recordings were submitted and uploaded to a secure cloud database, where they were accessed by our researchers. All participants had no diagnosed neurological or speech disorders, and all participants were nonsmokers. A total of 267 participants were included (170 men: 113 nondiabetic and 57 T2DM; and 97 women: 79 nondiabetic and 18 T2DM), recording a total of 18,465 voice samples.

Data Split

To evaluate voice changes due to T2DM status (and not to confounding factors such as age or BMI), the data was segmented by participant ID into an age-matched and BMI-matched dataset (referred to as the matched dataset") for both men and women. Because of a large proportion of nondiabetic participants that skewed younger than the T2DM sample, a number of nondiabetic recordings that had no age or BMI match in the T2DM arm were excluded from the matched sample. An equal number of T2DM and nondiabetic participants

Mayo Clin Proc Digital Health ■ December 2023;1(4):534-544 ■ https://doi.org/10.1016/j.mcpdig.2023.08.005
www.mcpdigitalhealth.org

were included in the matched dataset, such that 50% of men with T2DM and 61% of women with T2DM were included. Statistical analysis and prediction model training were performed on the matched dataset, and the remaining data was used to test the fully trained model (referred to as the test dataset). The increased proportion of women with T2DM was to allow for sufficient recordings for analysis whereas leaving sufficient data for testing the trained prediction model. In total, there were 4011 recordings and 1008 recordings for men and women, respectively, in the matched samples.

Feature Extraction

To allow for comparison with the most common features present in previous findings, voice features corresponding to pitch, intensity, harmonic noise ratio (HNR), shimmer, and jitter were extracted. Voice features were extracted using Parselmouth, a publicly available Python integration for Praat, a voice and speech analysis software.[27,28] A total of 14 voice features were extracted from each audio recording. Labels and descriptions of voice features can be found in Appendix A (available online at https://www.mcpdigitalhealth.org/).

Increased absolute values of shimmer and jitter are associated with increased perceived breathiness, hoarseness, and roughness in the voice, which can be linked to certain pathologies. These features are typically used exclusively in sustained phonation of vowel sounds; however, they have been found to be useful in identifying dysphonia when calculated from an entire sentence recording.[29] For this reason, jitter and shimmer values were chosen to be evaluated in addition to the pitch, intensity, and harmonic noise ratio vocal parameters.

Model

Logistic regression, Gaussian NB, and SVM were selected as models for the analysis, which are in line with models used in related work. Justification for these models and for exclusion of other models can be found in Appendix B.[30–36] A 5-fold cross-validation was performed on the matched dataset to find the optimal model, feature set, and threshold for prediction. The folds were segmented by individual and diabetic status, such that all recordings from an individual were placed into the same fold and there were the same number of nondiabetic and T2DM individuals in each fold. After cross-validation, the matched dataset was used to retrain the model. The model performance was assessed on the testing set (ie, the data not used in the matched dataset) using the retrained model, feature set, and threshold determined through cross-validation of the matched data. The procedure for the initial model experimentation is presented in Figure 1. Class prediction was performed using the T2DM probability results from the model prediction for each recording. If the probability of the recording was greater than or equal to a set threshold, the individual was predicted to T2DM, and if the probability model results were less than the threshold, the individual was predicted to be nondiabetic.

After the initial model implantation, there were a few methods applied to improve the prediction accuracy, the methods are as follows:

1. Average all prediction probabilities from an individual,
2. Average T2DM prevalence at the participant's age with prediction probability results,

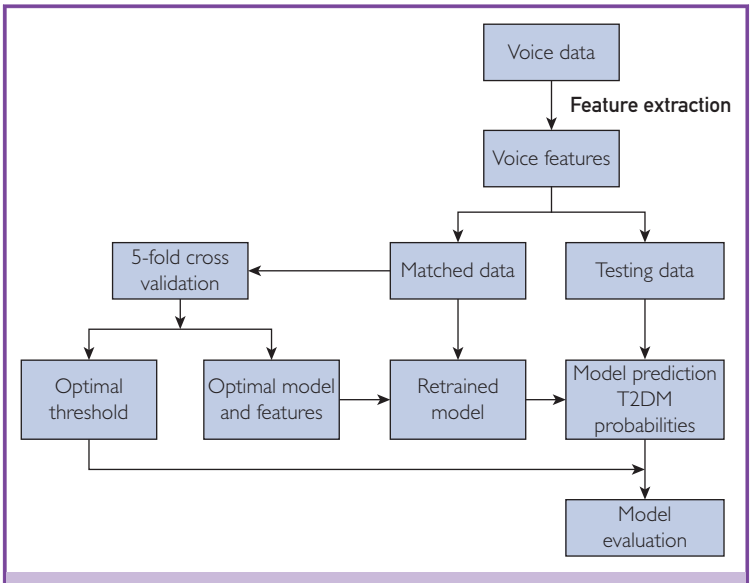

FIGURE 1. Flowchart of analysis and machine learning process. T2DM, type 2 diabetes mellitus.

TABLE 2. Acoustic Analysis of Audio Recordings.[a,b]

	Entire Dataset					Matched Dataset				
Feature	ND	T2DM	*P*	*P*-adj	Cohen's d	ND	T2DM	*P*	*P*-adj	Cohen's d
Woman										
Meanf0	213.23±34.49	191.73±37.38	<.001	<.001	0.6	210.03±37.22	194.65±37.13	<.001	<.001	0.41
stdevF0	38.66±19.97	32.88±19.63	<.001	<.001	.29	44.58±21.19	34.11±19.99	<.001	<.001	0.51
Meaninten	65.39±6.33	65.6±5.61	.3	>.99	.04	64.72±5.65	66.46±4.65	<.001	<.001	0.34
stdevInten	7.7±2.96	8.76±3.43	<.001	<.001	.33	8.43±3.03	8.59±3.26	.29	>0.99	0.05
HNR	12.1±3.95	11.53±4.32	<.001	<.001	.14	11.15±3.79	11.74±4.42	.004	.06	0.14
Localshimmer	0.12±0.04	0.12±0.03	.37	>.99	.03	0.12±0.04	0.12±0.04	<.001	<.001	0.2
Localdbshimmer	1.13±0.27	1.13±0.25	.5	>.99	0.02	1.19±0.27	1.13±0.27	<.001	<.001	0.24
apq3Shimmer	0.05±0.02	0.05±0.02	.002	.03	0.1	0.06±0.02	0.05±0.02	<.001	<.001	0.23
apq5Shimmer	0.07±0.03	0.07±0.03	.09	>.99	0.06	0.08±0.03	0.07±0.03	<.001	<.001	0.24
apq11Shimmer	0.11±0.04	0.11±0.03	.02	.25	0.08	0.12±0.04	0.11±0.04	<.001	<.001	0.2
Localjitter	0.02±0.01	0.02±0.01	.86	>.99	0.01	0.02±0.01	0.02±0.01	<.001	<.001	0.44
LocalabsoluteJitter	0.00010±0.00005	0.00012±0.00005	<.001	<.001	0.24	0.00012±0.00005	0.00011±0.00005	.03	.45	0.11
rapJitter	0.011±0.004	0.010±0.004	.02	.24	0.08	0.012±0.004	0.010±0.004	<.001	<.001	0.4
ppq5Jitter	0.011±0.004	0.011±0.004	.97	>.99	0.001	0.013±0.004	0.011±0.004	<.001	<.001	0.41
Man										
meanF0	140.04±29.71	140.2±32.79	.8	>.99	0.01	141.4±33.7	139.35±33.91	.06	.82	0.06
stdevF0	25.34±23.48	28.78±25.56	<.001	<.001	0.14	29.29±25.6	29.49±26.73	.81	>.99	0.01
meanInten	67.14±6.70	62.64±8.08	<.001	<.001	0.61	65.87±6.29	62.88±7.90	<.001	<.001	0.42
stdevInten	8.00±3.56	8.92±3.00	<.001	<.001	0.28	8.34±2.89	9.02±2.96	<.001	<.001	0.23
HNR	10.61±3.34	10.76±3.12	.02	.30	0.05	10.53±3.11	10.53±3.01	.95	>.99	0.002
localShimmer	0.12±0.03	0.13±0.03	<.001	<.001	0.19	0.12±0.03	0.13±0.03	<.001	<.001	0.15
localdbShimmer	1.16±0.23	1.21±0.22	<.001	<.001	0.25	1.18±0.22	1.22±0.21	<.001	<.001	0.17
apq3Shimmer	0.05±0.02	0.05±0.02	.02	.34	0.05	0.05±0.02	0.05±0.02	.52	>.99	0.02
apq5Shimmer	0.07±0.02	0.08±0.02	<.001	<.001	0.2	0.08±0.02	0.08±0.02	.003	.05	0.09
apq11Shimmer	0.12±0.04	0.14±0.04	<.001	<.001	0.35	0.13±0.04	0.14±0.04	<.001	<.001	0.28
localJitter	0.02±0.01	0.02±0.01	<.001	<.001	0.24	0.022±0.007	0.024±0.007	<.001	<.001	0.2
localabsoluteJitter	0.00016±0.00006	0.00018±0.00006	<.001	<.001	0.21	0.00017±0.00007	0.00018±0.00007	<.001	<.001	0.19
rapJitter	0.011±0.004	0.011±0.004	<.001	<.001	0.17	0.011±0.004	0.011±0.004	<.001	<.001	0.15
ppq5Jitter	0.012±0.004	0.013±0.004	<.001	<.001	0.22	0.012±0.004	0.013±0.004	<.001	<.001	0.22

[a]apq, amplitude perturbation quotient; F0, fundamental frequency; HNR, harmonic to noise ratio; ND, nondiabetic; rap, relative average perturbation; T2DM, type 2 diabetic mellitus.
[b]*P*-values from independent 2-tailed *t* test. Adjusted *P*-values (*P*-adj) from Bonferroni correction.

TABLE 3. Model Prediction Results.

Model ensemble	Type	Model	Threshold	5-fold CV				Testing			
				Sensitivity	Specificity	Accuracy	BCA	Sensitivity	Specificity	Accuracy	BCA
Woman											
Just voice	All recordings	3-feature LR	0.54	0.54±0.14	0.75±0.15	0.67±0.14	0.65±0.14	0.58	0.71	0.7	0.64
Just voice	Averaged per individual	3-feature LR	0.54	0.53±0.07	0.90±0.22	0.72±0.13	0.72±0.13	0.57	0.87	0.84	0.72
Voice and age	All recordings	3-feature LR	0.31	0.67±0.18	0.70±0.11	0.68±0.13	0.67±0.13	0.69	0.80	0.80	0.75
Voice and age	Averaged per individual	3-feature LR	0.3	0.63±0.37	0.83±0.21	0.73±0.28	0.73±0.28	0.71	0.75	0.75	0.73
Voice and BMI	All recordings	3-feature LR	0.31	0.69±0.18	0.71±0.24	0.70±0.19	0.70±0.19	0.73	0.71	0.71	0.72
Voice and BMI	Averaged per individual	3-feature LR	0.3	0.73±0.23	0.77±0.29	0.75±0.22	0.75±0.22	0.71	0.91	0.89	0.81
Voice, age, and BMI	All recordings	3-feature LR	0.24	0.70±0.22	0.72±0.19	0.71±0.19	0.71±0.19	0.79	0.83	0.82	0.81
Voice, age, and BMI	Averaged per individual	3-feature LR	0.24	0.63±0.37	0.83±0.21	0.73±0.28	0.73±0.28	0.71	0.93	0.91	0.82
Man											
Just voice	All recordings	2-feature NB	0.46	0.49±0.09	0.70±0.13	0.60±0.08	0.60±0.09	0.52	0.74	0.69	0.63
Just voice	Averaged per individual	2-feature NB	0.46	0.59±0.16	0.70±0.19	0.66±0.11	0.65±0.11	0.54	0.75	0.70	0.64
Voice and age	All recordings	2-feature NB	0.29	0.56±0.06	0.73±0.10	0.64±0.03	0.65±0.04	0.60	0.87	0.81	0.74
Voice and age	Averaged per individual	2-feature NB	0.28	0.58±0.19	0.73±0.24	0.66±0.13	0.65±0.13	0.82	0.87	0.86	0.85
Voice and BMI	All recordings	2-feature NB	0.28	0.58±0.08	0.75±0.11	0.66±0.03	0.66±0.03	0.56	0.75	0.71	0.65
Voice and BMI	Averaged per individual	2-feature NB	0.28	0.59±0.12	0.79±00.16	0.69±0.11	0.69±0.11	0.57	0.76	0.72	0.67
Voice, age, and BMI	All recordings	2-feature NB	0.22	0.58±0.06	0.74±0.11	0.67±0.03	0.66±0.03	0.66	0.84	0.8	0.75
Voice, age, and BMI	Averaged per individual	2-feature NB	0.22	0.69±0.15	0.73±0.13	0.70±0.10	0.71±0.11	0.75	0.89	0.86	0.82

BCA, balanced class accuracy; BMI, body mass index; LR, logistic regression; NB, Gaussian Naïve Bayes

Model prediction results from averaging and demographic characteristic addition methods. See Appendix B for averaging and demographic addition methods.

3. Average T2DM prevalence at the participant's BMI with prediction probability results, and
4. A combination of the above methods.

More details on feature selection, model implementation, optimal model selection, and ensembling methods can be found in Appendix B. All model training, testing, and validation were performed in Python (Python Software Foundation) using the Scikit-learn library (version 1.2.0, Python version 3.10.8).

Statistical Analyses

Student's independent 2-tailed *t* test was performed on demographic characteristic and vocal parameter values between the group with T2DM and the nondiabetic group. The Bonferroni correction was applied to *P*-values in the vocal parameter statistical analysis to account for multiple comparisons. Statistical analysis was performed in Python, using the SciPy package (version 1.9.3). Statistical significance is defined as $P<.05$.

Model accuracy was assessed on the basis of the accuracy, sensitivity, and specificity of the trained models. Equations for the calculation of these metrics can be found in Appendix B. The net reclassification index (NRI) was calculated to quantify the gain of adding vocal features to age and BMI risk factors for T2DM (see Appendix B).

RESULTS

The demographic characteristic information for the entire dataset and the age-matched and BMI-matched dataset can be found in Table 1. Fourteen voice features were extracted from each voice recording. Pitch, pitch SD, and apq3Shimmer were significant after adjustment between the nondiabetic and women in both the matched dataset and the entire dataset. For men, mean intensity, intensity SD, local shimmer, local absolute shimmer, apq11 shimmer, local jitter, local absolute jitter, RAP jitter, and 5-point percent perturbation quotient (ppq5) jitter were significant between the nondiabetic and diabetic individuals in both the matched dataset and the entire dataset (Table 2).

Model Formulation

A 5-fold cross-validation of the age-matched and BMI-matched dataset was used to evaluate model performance and identify the optimal threshold for model prediction probability results. Features were selected if they had an unadjusted $P<.05$ in both the entire dataset and the matched dataset and a Cohen's d >0.2 in the matched dataset (Table 2). In both men and women, there were 4 features that satisfied these criteria. The features used in the model evaluation were, in order of addition according to descending Cohen's d, pitch SD, mean pitch, RAP jitter, and apq3 shimmer for women, and mean intensity, apq11 shimmer, intensity SD, and ppq5 jitter for men. Model testing was performed on the test dataset, using the same optimal model and threshold as determined by cross-validation.

Female Model Results. The optimal model for women after cross-validation was a 3-feature LR model and had a specificity of 0.75 ± 0.15, a sensitivity of 0.54 ± 0.14, and an optimal threshold of 0.54 for all data recordings (Table 3, Appendix C). The 3 features used in the female model were the mean pitch, pitch SD, and RAP jitter. If the model prediction probabilities of T2DM for all recordings for a participant were averaged, the 3-feature LR female model had a specificity of 0.90 ± 0.22 and a sensitivity of 0.53 ± 0.07 for the threshold of 0.54. The female test set had a final specificity of 0.71 and sensitivity of 0.58 (3-feature LR model for single recordings), and if prediction probabilities for all recordings for a participant were averaged, the 3-feature LR model had a specificity of 0.87 and sensitivity of 0.57 (Table 3).

Male Model Results. The optimal model for males was a 2-feature NB and had a specificity of 0.70 ± 0.13, a sensitivity of 0.49 ± 0.09, and an optimal threshold of 0.46 for all data recordings (Table 3, Appendix C). The features used in the male model were the mean intensity and apq11 shimmer values. If the model prediction probabilities of T2DM for all recordings for each participant were averaged, the 2-feature NB male model had a specificity of 0.70 ± 0.19 and a sensitivity of 0.59 ± 0.16

Mayo Clin Proc Digital Health ■ December 2023;1(4):534-544 ■ https://doi.org/10.1016/j.mcpdig.2023.08.005
www.mcpdigitalhealth.org

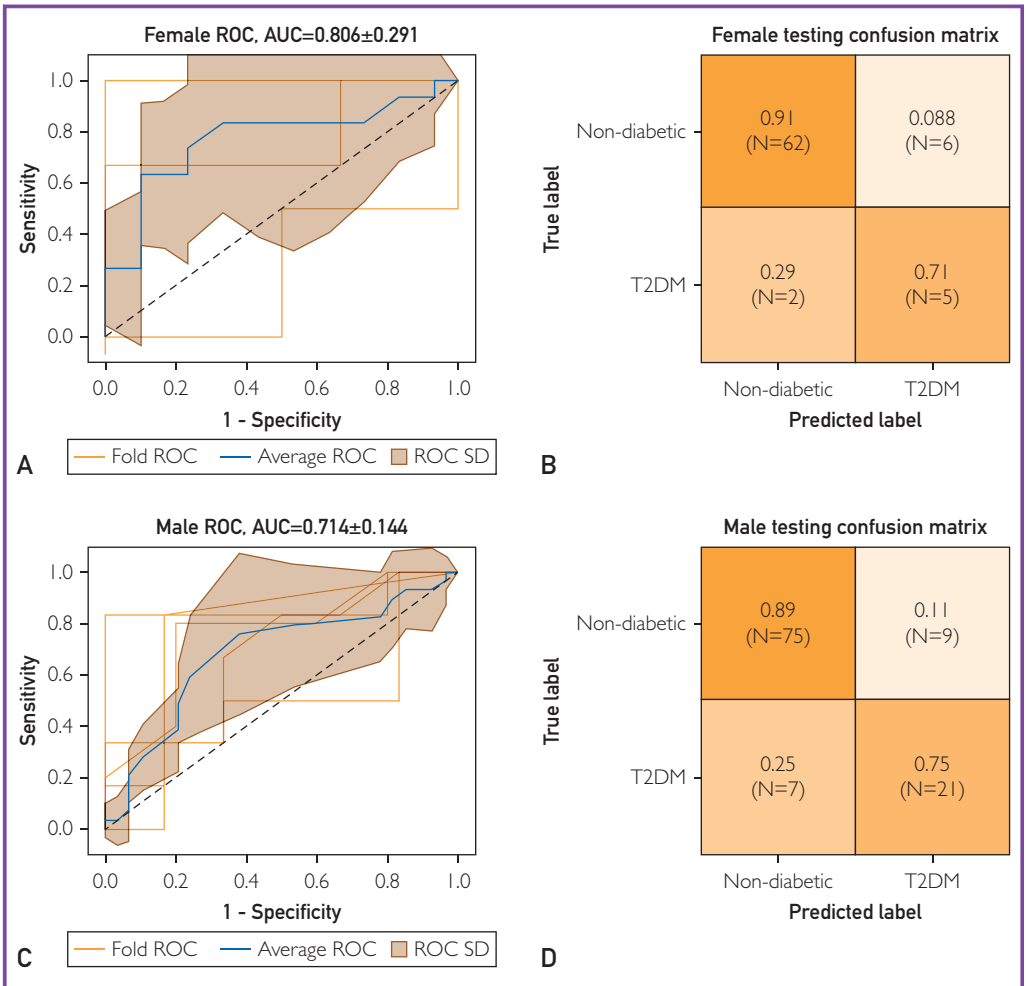

FIGURE 2. Results from optimal model. (A) ROC curve for cross-validation of the matched dataset for women (B) Confusion matrix for the test set for women (C) ROC curve for cross-validation of the matched dataset for men (D) Confusion matrix for the test set for men. N is the number of participants. T2DM, type 2 diabetes mellitus; ROC, receiver operator curve; AUC, area under the ROC.

with a 0.48 threshold (Table 3). The male test set had a final specificity of 0.74 and sensitivity of 0.52 when predicting individual recordings (2-feature NB model), and if all the prediction results for an individual were averaged, the 2-feature NB model had a specificity of 0.75 and sensitivity of 0.54 (Table 3).

Incorporation of Demographic Characteristic Data

To increase model accuracy, age and BMI prevalence were incorporated into the prediction methodology. The features and model type were kept the same as the optimal model determined in the voice-only prediction, so the 3-feature LR model was used for women and the 2-feature NB model was used for men.

Female Ensemble Results. The optimal ensemble model prediction was achieved by averaging the female voice recording results with the BMI prevalence of T2DM. The model had an optimal accuracy of 0.75±0.22, with a specificity of 0.77±0.29 and a sensitivity of 0.73±0.23 from cross-validation of the matched dataset, and had an accuracy of 0.89, specificity of 0.91, and sensitivity of 0.71 when predicting the test set (Table 3, Figure 2). Looking at the receiver operator

curve (Figure 2), the average area under the curve (AUC) was 0.81±0.29 for cross-validation. The folds consisted of 3 folds with perfect prediction (AUC=1), 1 fold with moderate prediction results, and 1 fold with poor prediction results (AUC<0.5), indicating some variability in the model prediction results.

Male Ensemble Results. The optimal ensemble model prediction was obtained by averaging the male voice recording prediction results with the age and BMI prevalence of T2DM. This model had an optimal accuracy of 0.70±0.10, with a specificity of 0.73±0.13 and a sensitivity of 0.69±0.15 from cross-validation of the matched dataset, and had an accuracy of 0.86, a specificity of 0.89, and a sensitivity of 0.75 when predicting the test set (Table 3, Figure 2). The average AUC was 0.71±0.14 for cross-validation (Figure 2). Overall, there was less variability in the male cross-validation results, particularly for high sensitivity values, in which 4/5 folds fell within the ROC SD.

Net Reclassification Index Calculation

Net reclassification index was calculated for both men and women in the matched dataset. For women, the overall NRI from adding voice to age and BMI was 0.36, and for men, the overall NRI was 0.28. Details of NRI calculations can be found in Appendix D.

DISCUSSION

Overall, we found distinct differences between the voices of individuals with and without T2DM. These differences vary between men and women and reinforce previous findings that the vocal manifestations of diabetes are sex-specific. The most accurate prediction method involved an ensemble model with T2DM prevalence at the participant's age and BMI for men and the participant's BMI for women, resulting in a maximum test accuracy of 0.89 for women and 0.86 for men. Optimal models were a 2-vocal-feature NB implementation for men and a 3-vocal-feature LR for women.

Different features were used in the prediction between men and women. In women, the predictive features were mean pitch, pitch SD, and RAP jitter, and in men, mean intensity and apq11 shimmer were used. In simple terms, the variation in these features found that women with T2DM reported a slightly lower pitch with less variation, and men with T2DM reported slightly weaker voices with more variation. These differences likely stem from differences in disease symptom manifestations between the sexes. For example, muscle weakness and atrophy, which have been linked to vocal weakness and instability,[7] occur in T2DM and are more common in men with T2DM than in women with T2DM.[37] On contrary, women with T2DM are more likely to experience high extracellular water content and edema.[38] Swelling and edema of the vocal cords reduce the pitch and vibratory characteristics, resulting in a parameter decrease similar to what was seen in our results.[39] Laryngoscopy and visualization of the vocal cords in T2DM should be performed in future studies to confirm these findings. Furthermore, cognitive function decline and major depressive disorder (MDD) occur at a higher prevalence in women with T2DM than in men with T2DM ,[9,40] and peripheral neuropathy occurs at a higher prevalence in men with T2DM .[41] Cognitive impairment has been shown to have a significant effect on the voice with strong predictive capabilities,[10] and MDD has been linked to voice changes such as slower speech and a lower pitch.[13] Sex differences in T2DM have become increasingly prominent, as seen in the contrasting predictive features, and future research should carefully account for this for a more comprehensive insight.

Comparing the results to previous findings, fundamental frequency in women with T2DM has been shown to decrease in previous work,[14] which aligns with the results presented here. In men with T2DM , previous findings varied from no significant features to all features being significant, except jitter.[14,16,25] Indeed, there were fewer features used in male prediction than in female prediction, and features used in prediction did not include pitch, jitter, or HNR. Furthermore, previous work primarily used a stand-alone microphone or specific recording device to collect data rather than a smartphone and an application-based approach. Application-based recording offers considerable advantages over stand-alone microphones for voice data collection. Its accessibility options allow

researchers to capture data using widely available devices, expanding participant inclusivity. In addition, smartphone recording has the potential to capture real-world situations, such as speech and interactions in familiar surroundings. Overall, researchers can obtain insights into how voices change in everyday scenarios.

Incorporating vocal parameters, age, and BMI into an ensemble prediction model was able to achieve over 70% accuracy in an age-matched and BMI-matched dataset and achieved even higher accuracy in the unmatched test dataset. This result indicates that even the simple incorporation of age and BMI into an ensemble model with voice creates an accurate prediction methodology for T2DM. Furthermore, adding vocal features to demographic characteristic information like age or BMI resulted in a positive NRI for both men (NRI=0.28) and women (NRI=0.36), indicating that voice combined with age and BMI performs better at classifying risk than age and BMI alone.

Finally, there were some limitations to the presented methodology. The duration of T2DM may influence the voice,[25] and future work should incorporate the collection of T2DM duration into the study protocol, and cognitive function metrics and MDD status, to account for additional factors that may influence the voice. In addition, only a crude ensemble of demographic characteristic and vocal features was used in the final model implementation. Future work could explore alternative ways to incorporate demographic data into the model results and incorporate additional demographic characteristic features. Finally, although there were thousands of audio recordings, the female T2DM population was small in the presented work, resulting in some prediction variability in the model. Future work should include a larger cohort of individuals to confirm the findings presented here.

CONCLUSION

The material presented here reports a promising application of voice analysis for T2DM detection. Although the results are encouraging, further research with larger and more diverse cohorts is required to validate its effectiveness and generalizability. Nevertheless, our findings highlight the potential of voice analysis as an accessible and cost-effective screening tool. An implementation of voice assessment could aid in early intervention and management of T2DM, and continued development could reduce the rising burden of the disease and improve health care outcomes.

POTENTIAL COMPETING INTERESTS

J.K., A.T., and Y.F. are all employees of the source of funding for the study, Klick Inc. Y.F. is listed as an inventor on 2 patents for the estimation of blood glucose using voice (patent numbers WO2022109713A1 and WO2022109714A1).

ACKNOWLEDGMENTS

The authors would like to thank everyone within Klick Applied Sciences for their support, particularly Jouhyun Jeon for her contribution to the study design and consultation during data analysis.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **ADA,** American Diabetes Association; **APQ,** amplitude of the perturbation quotient; **BMI,** body mass index; **HNR,** harmonic noise ratio; **LR,** logistic regression; **MDD,** major depressive disorder; **NB,** Naïve Bayes; **NRI,** net reclassification index; **RAP,** relative average perturbation; **SVM,** Support Vector Machine; **T2DM,** type 2 diabetes mellitus

Grant Support: The study was internally funded by Klick Inc. The study sponsor did not have a role in study design, collection, analysis, and interpretation of data, writing of the report, or decision to submit the paper for publication.

Correspondence: Jaycee M. Kaufman, MSc, 175 Bloor St E Suite 300, Toronto, ON, Canada M4W 3R8 (jmorgankaufman@klick.com).

ORCID

Jaycee M. Kaufman: https://orcid.org/0000-0001-8183-0206Yan Fossat: https://orcid.org/0000-0002-1271-2633

REFERENCES

1. Bommer C, Sagalova V, Heesemann E, et al. Global economic burden of diabetes in adults: projections from 2015 to 2030. *Diabetes Care*. 2018;41(5):963-970. https://doi.org/10.2337/dc17-1962.

2. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. *Diabetologia*. 2019;62(1):3-16.
3. Zhang Z. Mechanics of human voice production and control. *J Acoust Soc Am*. 2016;140(4):2614.
4. Sidorova J, Carbonell P, Čukić M. Blood glucose estimation from voice: first review of successes and challenges. *J Voice*. 2022;36(5):737.e1-737.e10.
5. Yagihashi S, Mizukami H, Sugimoto K. Mechanism of diabetic neuropathy: where are we now and where to go? *J Diabetes Investig*. 2011;2(1):18-32.
6. Ciarambino T, Crispino P, Leto G, Mastrolorenzo E, Para O, Giordano M. Influence of gender in diabetes mellitus and its complication. *Int J Mol Sci*. 2022;23(16):8850.
7. Weinreb SF, Piersiala K, Hillel AT, Akst LM, Best SR. Dysphonia and dysphagia as early manifestations of autoimmune inflammatory myopathy. *Am J Otolaryngology*. 2021;42(1):102747.
8. Hamdan AL, Kurban Z, Azar ST. Prevalence of phonatory symptoms in patients with type 2 diabetes mellitus. *Acta Diabetol*. 2013;50(5):731-736.
9. Palomo-Osuna J, Failde I, De Sola H, Dueñas M. Differences in cognitive function in women and men with diabetic peripheral neuropathy with or without pain. *Int J Environ Res Public Health*. 2022;19(24):17102.
10. López-de-Ipiña K, Martinez-de-Lizarduy U, Calvo PM, et al. On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment. *Neural Comput Appl*. 2020; 32(20):15761-15769.
11. Kato S, Homma A, Sakuma T. Easy screening for mild Alzheimer's disease and mild cognitive impairment from elderly speech. *Curr Alzheimer Res*. 2018;15(2):104-110.
12. Zhao Q, Fan HZ, Li YL, et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. *Front Psychiatry*. 2022:13.
13. Wang J, Zhang L, Liu T, Pan W, Hu B, Zhu T. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC Psychiatry*. 2019;19:300.
14. Pinyopodjanard S, Suppakitjanusant P, Lomprew P, Kasemkosin N, Chailurkit L, Ongphiphadhanakul B. Instrumental acoustic voice characteristics in adults with type 2 diabetes. *J Voice*. 2021;35(1):116-121.
15. Gölaç H, Atalik G, Türkcan AK, Yilmaz M. Disease related changes in vocal parameters of patients with type 2 diabetes mellitus. *Logoped Phoniat Vocol*. 2022;47(3):202-208.
16. Chitkara D, Sharma R. Voice based detection of type 2 diabetes mellitus. In: *2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE publications; 2016:83-87.
17. Sara JDS, Maor E, Orbelo D, Gulati R, Lerman LO, Lerman A. Noninvasive voice biomarker is associated with incident coronary artery disease events at follow-up. Vol 97. Elsevier. *Mayo Clin Proc*. 2022;95(5):835-846.
18. Alam MZ, Simonetti A, Brillantino R, et al. Predicting pulmonary function from the analysis of voice: a machine learning approach. *Front Digit Health*. 2022;4:750226.
19. Costantini G, Dr VC, Robotti C, et al. Deep learning and machine learning-based voice analysis for the detection of COVID-19: a proposal and comparison of architectures. *Knowl-Based Syst*. 2022;253:109539.
20. Costantini G, Cesarini V, Di Leo P, et al. Artificial intelligence-based voice assessment of patients with Parkinson's disease off and on treatment: machine vs. deep-learning comparison. *Sensors (Basel)*. 2023;23(4):2293.
21. Albadr MAA, Tiun S, Ayob M, Al-Dhief FT. Particle swarm optimization-based extreme learning machine for Covid-19 detection. *Cognit Comput*. 2022:1-16.
22. AL-Dhief FT, Latiff NMA, Baki MM, Malik NNNA, Sabri N, Albadr MAA. Voice pathology detection using support vector machine based on different number of voice signals. In: 2021 26th IEEE Asia-Pacific Conference on Communications (APCC). IEEE publications; 2021:1-6.
23. Syed S, Rashid M, Hussain S, Imtiaz A, Abid H, Zahid H. Inter classifier comparison to detect voice pathologies. *Math Biosci Eng*. 2021;18(3):2258-2273.
24. Suppakitjanusant P, Kasemkosin N, Sivapiromrat AK, et al. Predicting glycemic control status and high blood glucose levels through voice characteristic analysis in patients with cystic fibrosis-related diabetes (CFRD). *Sci Rep*. 2023;13(1): 8617.
25. Hamdan AL, Jabbour J, Nassar J, Dahouk I, Azar ST. Vocal characteristics in patients with type 2 diabetes mellitus. *Eur Arch Otorhinolaryngol*.. 2012;269:1489-1495.
26. American Diabetes Association Professional Practice Committee. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. *Diabetes Care*. 2022;45(suppl 1):S17-S38. https://doi.org/10.2337/dc22-S002.
27. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: A Python interface to Praat. *J Phon*. 2018;71:1-15. https://doi.org/10.1016/j.wocn.2018.07.001.
28. Boersma P, Weenink D. *Praat, a System for Doing Phonetics by Computer, version 3.4*. Institute of Phonetic Sciences of the University of Amsterdam; 1996.
29. Ancillao A, Galli M, Mignano M, Dellavalle R, Albertini G. Quantitative analysis of pathological female human voice by processing complete sentences recordings. *J Laryngol Voice*. 2013;3(2): 46.
30. Althnian A, AlSaeed D, Al-Baity H, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Appl Sc*. 2021;11(2):796.
31. India State-Level Disease Burden Initiative Diabetes Collaborators. The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990–2016. *Lancet Glob Health*. 2018;6(12):E1352-E1362.
32. Daousi C, Casson IF, Gill GV, MacFarlane IA, Wilding JP, Pinkney JH. Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. *Postgrad Med J*. 2006;82(966):280-284. https://doi.org/10.1136/pmj.2005.039032.
33. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiol (Camb, Mass)*. 2014;25(1): 114-121.
34. White T, Algeri S. Estimating the lifetime risk of a false positive screening test result. *Plos one*. 2023;18(2):e0281153.
35. Herman C. What makes a screening exam good? *Virtual Mentor*. 2006;8(1):34-37.
36. Kaur G, Lakshmi PVM, Rastogi A, et al. Diagnostic accuracy of tests for type 2 diabetes and prediabetes: A systematic review and meta-analysis. *PloS One*. 2020;15(11):e0242415.
37. Sayer AA, Dennison EM, Syddall HE, Gilbody HJ, Phillips DIW, Cooper C. Type 2 diabetes, muscle strength, and impaired physical function? *Diabetes Care*. 2005;28(10): 2541-2542.
38. Low S, Ng TP, Lim CL, et al. Higher ratio of extracellular water to total body water was associated with reduced cognitive function in type 2 diabetes. *J Diabetes*. 2021;13(3):222-231.
39. Dewan K, Chhetri DK, Hoffman H. Reinke's edema management and voice outcomes. *Laryngoscope Investig Otolaryngol*. 2022;7(4):1042-1050.
40. Deischinger C, Dervic E, Leutner M, et al. Diabetes mellitus is associated with a higher risk for major depressive disorder in women than in men. *BMJ Open Diabetes Res Care*. 2020;8(1): e001430.
41. Hicks CW, Selvin E. Epidemiology of peripheral neuropathy and lower extremity disease in diabetes. *Curr Diab Rep*. 2019; 19(10):86.