

**Ministry of Higher Education and  
Scientific Research  
University of Mosul  
College of Computer Science and  
Mathematics  
Department of Computer Science**



# **Automatic Smokers Classification from Speech Signal Using Machine Learning**

**A Thesis Submitted to the Council of the College of  
Computer Science and Mathematics  
University of Mosul  
as a Partial Fulfillment of Requirements  
for the Degree of Master of Science  
in  
Computer Science**

**By**

**Ibrahim Khudhur Zaal Majnan**

**Supervised by**

**Assist. Prof. Dr. Yusra Faisal Mohammad Al-Irhayim**

---

**2024 A.D.**

**1445 A.H.**

## **Abstract**

The human voice is affected and its quality decreases significantly when smoking. Some habits that a person practices have a very clear impact on their health. The most important of these habits are smoking, alcohol, and drug use, and there is a big difference between the voice of smokers and non-smokers.

Therefore, a system was designed to classify smokers from non-smokers and their gender based on the speech signal using machine learning algorithms using python language. This study focuses on the automatic classification of smokers using speech signals.

The suggested system uses 200 speech signals from 20 smokers and non-smokers, 10 male and 10 female. Ages range from 18 to 30. The dataset was manually recorded with a Zoom F1 microphone. Islamic greeting is the sentence recorded: (السلام عليكم ورحمة الله وبركاته) in the audio signal. Everyone repeated it 10 times.

After pre-processing the dataset and analyzing the signals to extract features, two types of features are extracted: Mel Frequency Cepstral Coefficients (MFCC) and acoustic features (F0, jitter, shimmer, and HNR) extract from signals. The back-propagation neural network, support vector machine (SVM), and decision tree (DT) cascade model start training after pre-processing the dataset. Data classification is a key use of machine learning algorithms.

The performance of the models was evaluated using the confusion matrix algorithm, which the back-propagation network model achieved (precision = 0.90 for male and precision = 0.73 for female) the SVM

model achieved (precision = 0.94 for male precision = 0.89 for female), and the Decision Tree (DT) model achieved (precision = 0.87 for male and precision = 0.72 for female). The SVM algorithm has obtained the classification rate, and it classifies 4 people out of 5 people correctly.



وزارة التعليم العالي والبحث العلمي  
جامعة الموصل  
كلية علوم الحاسوب والرياضيات  
قسم علوم الحاسوب

## تصنيف تلقائي للمدخنين من إشارة الكلام باستخدام التعلم الآلي

رسالة مقدمة  
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل  
كجزء من متطلبات نيل شهادة ماجستير علوم في  
علوم الحاسوب

من قبل

ابراهيم خضر زعل مجنان

بإشراف  
أ.م.د. يسرى فيصل محمد الرحيم

## الخلاصة

يتأثر صوت الإنسان، وتنخفض جودته بشكل ملحوظ عند التدخين. العادات التي يمارسها الإنسان لها تأثير واضح جداً على صحته. ومن أهم هذه العادات التدخين والكحول وتعاطي المخدرات، وهناك فرق كبير بين أصوات المدخنين وغير المدخنين.

تم تصميم نظام لتصنيف المدخنين من غير المدخنين وجنسهم بناء على إشارة الكلام باستخدام خوارزميات التعلم الآلي في لغة بايثون. تركز هذه الدراسة على التصنيف التلقائي للمدخنين باستخدام إشارات الكلام.

يستخدم النظام المقترح 200 إشارة كلامية من 20 مدخناً وغير مدخن، 10 ذكور و10 إناث. تتراوح الأعمار من 18 إلى 30 عاماً. وتم تسجيل مجموعة البيانات يدوياً باستخدام ميكروفون Zoom F1. التحية الإسلامية هي الجملة المسجلة: (السلام عليكم ورحمة الله وبركاته) في الإشارة الصوتية. كررها الأشخاص 10 مرات.

بعد المعالجة المسبقة لمجموعة البيانات وتحليل الإشارات لاستخراج الميزات، يتم استخراج نوعين من الميزات: معاملات ميل التردد Cepstral (MFCC) وميزات الطاقة (F0)، والارتعاش، والوميض، و (HNR) المستخرجة من الإشارات. تبدأ الشبكة العصبية ذات الانتشار الخلفي وجهاز ناقل الدعم (SVM) ونموذج التتالي لشجرة القرار (DT) في التدريب بعد المعالجة المسبقة لمجموعة البيانات. يعد تصنيف البيانات أحد الاستخدامات الرئيسية لخوارزميات التعلم الآلي.

تم تقييم أداء النماذج باستخدام خوارزمية مصفوفة الارتباك، والتي حققها نموذج شبكة الانتشار الخلفي (الدقة = 0.90 للذكور والدقة = 0.73 للإناث) حقق نموذج SVM (الدقة = 0.94 لدقة الذكور = 0.89 للإناث)، وتم تحقيق نموذج شجرة القرار (DT) (الدقة = 0.87 للذكور والدقة = 0.72 للإناث). حققت خوارزمية SVM أعلى معدل تصنيف وهو تصنيف 4 اشخاص من اصل 5 اشخاص بشكل صحيح.