



الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

الجمهورية العربية السورية
الجامعة السورية الخاصة
كلية الهندسة المعلوماتية
المشروع تخرج واحد

تحليل الصوت للكشف عن مرض باركنسون باستخدام خوارزميات التعلم الآلي

Voice Analysis to Detect Parkinson's Disease Using Machine Learning Algorithms

إعداد الطلاب:

ياسمين سيد أحمد

سحر محمد مروان الحايك

إشراف:

م. أنس عبد العزيز

د. مهيب النكري

2023 □ 2024

شهادة مشرف :

اسم..... :

تاريخ :

التوقيع:

الملخص

يقوم هذا المشروع بالكشف المبكر عن مرض باركنسون باستخدام تقنيات التعلم الآلي. يستند المشروع على تحليل تسجيلات صوتية للأفراد باستخدام نموذج التعلم الآلي يستخدم مميزات صوتية مثل التردد والإهتزازات والشدة. يهدف المشروع إلى تطوير أداة فعالة للكشف المبكر عن علامات محتملة لمرض باركنسون بدقة عالية. تم استخدام مجموعة بيانات تتألف من تسجيلات صوتية لأفراد مصابين وأصحاء، وتم تدريب ومقارنة خوارزميات مختلفة وتم الحصول على نتائج جيدة من الخوارزميات والمقارنة بينها، وذلك بعد تطبيق مراحل معالجة مختلفة على البيانات قبل التدريب. وأظهرت النتائج أن خوارزمية Random Forest حققت أعلى دقة بنسبة 96.6% ومعدل F1_Score بنسبة 96.7%. يشير ذلك إلى فعالية هذا النموذج في التمييز بين الأفراد المصابين والأصحاء، مما يبرز أهمية استخدام تقنيات الذكاء الاصطناعي في تحسين عمليات الكشف المبكر عن مرض باركنسون.

Abstract

This project focuses on early detection of Parkinson's disease using machine learning techniques. The project relies on analyzing audio recordings of individuals using a machine learning model that utilizes voice features such as frequency, vibrations, and intensity. The goal of the project is to develop an effective tool for early detection of potential signs of Parkinson's disease with high accuracy. A dataset comprising audio recordings of both affected and healthy individuals was used, and various algorithms were trained and compared. Good results were obtained from the algorithms after applying different data processing stages before training. The results showed that the Random Forest algorithm achieved the highest accuracy at 96.6% and an F1_Score of 96.7%. This indicates the effectiveness of this model in distinguishing between affected and healthy individuals, highlighting the importance of using artificial intelligence techniques to enhance early detection processes for Parkinson's disease.

جدول المحتويات

3.....	الملخص
4.....	Abstract
10.....	1 الفصل الأول: مقدمة
11.....	1.1 مقدمة عن المشروع
12.....	1.2 المشكلة التي يقوم المشروع بحلّها
13.....	1.3 الهدف من المشروع
14.....	1.4 مقدمة عن مرض باركنسون
19.....	1.5 استخدام أدوات الذكاء الصناعي في مرض باركنسون
22.....	1.6 التعلم الآلي machine learning
24.....	1.7 خوارزميات التعلم الآلي شائعة الاستخدام
24.....	1.7.1 الإنحدار اللوجستي logistic regression
26.....	1.7.2 شجرة القرار
27.....	1.7.3 خوارزمية SVM

28	1.7.4 خوارزمية Naïve Bayes
29	1.7.5 خوارزمية KNN
30	1.7.6 الغابة العشوائية
31	2 الفصل الثاني: الدراسة المرجعية
34	2.1 المنهجية الأولى
34	2.2 المنهجية الثانية
35	2.3 المنهجية الثالثة
37	3 الفصل الثالث: مجموعة البيانات ومعايير التقييم
38	3.1 مجموعة البيانات dataset
40	3.2 معايير التقييم Evaluation matrices
45	4 الفصل الرابع: تنفيذ المشروع
46	4.1 البيئة والأدوات المستخدمة
47	4.2 المنهجية
49	4.2.1 معالجة البيانات Data Preprocessing
50	4.2.2 اختيار المميزات
51	4.2.3 تقسيم البيانات إلى تدريب واختبار

52.....	4.2.4 تطبيق خوارزميات التعلم الآلي
53.....	5 الفصل الخامس: الخلاصة والنتائج
54.....	5.1 النتائج
58.....	5.2 الخلاصة
59.....	5.3 الأعمال المستقبلية
60.....	المراجع:

فهرس الاشكال :

الشكل 1 الهيكل العام لشجرة القرار DT (Guia et al., 2019)	27
الشكل 2 عمل خوارزمية Random Forest (Breiman, 2001)	30
الشكل 3 المخطط الصندوقي لمراحل عمل الدراسة المرجعية (Govindu & Palwe, 2023)	33
الشكل 4 مصفوفة الارتباك (confusion matrix) (Vujović, 2021)	42
الشكل 5 المنهجية المتبعة في المشروع	48
الشكل 6 البيانات بعد أن أصبحت متوازنة	50
الشكل 7 عينة من المميزات المختارة	51
الشكل 8 دقة خوارزمية RF	55
الشكل 9 خطأ خوارزمية RF على بيانات التدريب والاختبار	55
الشكل 10 مصفوفة confusion matrix لنتائج RF	56
الشكل 11 المقارنة بالدقة بين مختلف الخوارزميات	57
الشكل 12 المقارنة بالخطأ بين مختلف الخوارزميات	58

فهرس الجداول:

الجدول 1 وصف مميزات مجموعة البيانات (Govindu & Palwe, 2023)	38
الجدول 2 نتائج تطبيق الخوارزميات على بيانات الاختبار	54

1 الفصل الأول: مقدمة

1.1 مقدمة عن المشروع

في ظل التطور السريع للتكنولوجيا وتقدم مجالات التعلم الآلي ومعالجة اللغات الطبيعية، تتسارع الجهود نحو تحسين الخدمات الطبية والتشخيص المبكر للأمراض. يُعد مرض باركنسون، الذي يؤثر على جهاز الحركة ويتسبب في تغيرات في الصوت والحديث، تحدياً كبيراً يتطلب اهتماماً فورياً لتوفير العلاج الملائم. يتّسم هذا المشروع بمحاولة توظيف تقنيات التعلم الآلي لتحليل التسجيلات الصوتية بهدف الكشف المبكر عن علامات مرض باركنسون، مما يعزز من إمكانية التدخل الطبي الفعّال ويحسن جودة حياة المصابين. من خلال هذا النهج، نسعى إلى دمج الابتكار التكنولوجي بشكل مبتكر مع الرعاية الصحية لتحسين تشخيص ومعالجة مرضى باركنسون.

يأتي هذا المشروع كخطوة جديدة نحو تحسين عمليات الكشف المبكر عن مرض باركنسون باستخدام تقنيات التعلم الآلي. يقوم النظام المقترح بتحليل بيانات لتسجيلات صوتية من مرضى باركنسون وأفراد أصحاء حيث أنّ هذه البيانات تحوي مميزات صوتية تكون مؤشراً على وجود المرض. يهدف المشروع إلى توفير أداة فعّالة ومبتكرة للأطباء والمختصين لتسهيل عمليات الكشف المبكر والتدخل السريع.

1.2 المشكلة التي يقوم المشروع بحلّها

التحديات التي تواجه عمليات الكشف المبكر تتمثل في أن الطرق التقليدية قد تكون مكلفة وتتطلب وقتاً طويلاً. هذا يعني أن العديد من الأفراد الذين قد يعانون من أعراض بداية مرض باركنسون قد لا يحصلون على التشخيص المبكر الذي يمكن أن يساعد في بدء العلاج وتحسين نوعية حياتهم.

لحل هذه المشكلة، يقوم المشروع بالاعتماد على تقنيات تعلم الآلة ومعالجة اللغات الطبيعية لتحليل مجموعة من تسجيلات الصوت للأفراد، سواء كانوا مرضى باركنسون أو أفراداً أصحاء. يتمثل الهدف في تطوير نموذج تعلم آلي يستند إلى مميزات صوتية تستخرج من هذه التسجيلات، مثل التردد الصوتي، والإهتزازات، والتغيرات في الشدة.

باستخدام هذا النموذج، يمكن تحقيق كشف مبكر لعلامات محتملة لمرض باركنسون بشكل أكثر دقة وفعالية. هذا يمكن أن يساهم في تسهيل عمليات الكشف المبكر وتحسين جودة الرعاية الصحية المقدمة للأفراد المصابين بمرض باركنسون. وعلى الرغم من وجود العديد من الأبحاث في هذا المجال والعديد من المشاريع إلا أن تسليط الضوء على هذه القضية واستخدام خوارزميات مختلفة فيها والمقارنة بينها يساهم في تطوير المعرفة في هذا المجال.

1.3 الهدف من المشروع

المشروع الذي يستخدم بيانات تسجيلات صوتية للكشف عن مرض باركنسون يعكس أهمية توظيف التقنيات الحديثة في مجال الرعاية الصحية. يُعتبر الكشف المبكر عن هذا المرض أمراً حيوياً، حيث يمكن أن يؤدي إلى بداية العلاج في وقتٍ مبكر، مما يساهم في تحسين نوعية حياة المرضى وتأثيرات المرض على وظائفهم الحياتية.

يساعد استخدام بيانات التسجيل الصوتي على تحليل السمات الصوتية للتمييز بين التغيرات الصوتية التي تحدث نتيجة لمرض باركنسون والصوت الطبيعي. فإن هدفنا هو تطوير نماذج تعلم آلي قادرة على التمييز بين حالات المرض والأفراد الأصحاء باستخدام هذه البيانات، والمقارنة بين هذه الخوارزميات حسب معايير التقييم المناسبة لتحديد الأفضل بينها. بالتأكيد هذا يوفر أداة فعالة للكشف المبكر عن مرض باركنسون. ليس هذا فقط يمكن أن يقلل من التأخير في التشخيص، ولكنه أيضاً يفتح الباب للتدخل الطبي الفعال، مما يقلل من تقدم المرض ويحسن تجربة المرضى.

1.4 مقدمة عن مرض باركنسون

مرض باركنسون يُعتبر واحدًا من الأمراض العصبية المزمنة التي تؤثر على جهاز الحركة. يتميز هذا المرض بفقدان التحكم التدريجي في الحركة وتدهور القدرة على التنسيق بين العضلات، مما يؤدي إلى ظهور أعراض تتضمن الرعاش، الصعوبة في التوازن، والتصلب العضلي (Porter & Gupta, 2023).

مرض باركنسون هو اضطراب عصبي يتسم بتدهور نوعي في خلايا الدماغ المسؤولة عن إفراز مادة الدوبامين. هذا التدهور يؤدي إلى تأثير على الحركة الطبيعية للجسم، ويظهر بشكل أكثر شيوعًا عند كبار السن.

وأيضاً تعريف آخر لمرض باركنسون: هو اضطراب عصبي مزمن يؤثر على الحركة، وغالباً ما يتميز بالرجفة. يحدث نتيجة تلف الخلايا العصبية في جزء من الدماغ يُعرف بالمادة السوداء. هذه الخلايا العصبية المتلفة تتسبب في نقص الدوبامين، وهو ناقل عصبي يلعب دوراً مهماً في التحكم في الحركة (Sveinbjornsdottir, 2016).

يأتي دور الكشف المبكر عن مرض باركنسون بأهمية خاصة في توفير فرصة للتدخل الطبي المبكر وتحسين نوعية حياة المصابين (Govindu & Palwe, 2023). تمكين الأطباء من تشخيص المرض في مراحله الأولى يساعد في تقديم العلاج المناسب وإدارة الأعراض بفعالية، مما يعزز من فرص التحسين والاستقرار على المدى الطويل. والكشف المبكر عن مرض باركنسون أيضاً يمكن أن يساعد في إدارة الأعراض بشكل أكثر فعالية ويمكن أن يحسن جودة الحياة للمرضى. الكشف المبكر يمكن أن يساعد أيضاً في فهم تقدم المرض والتخطيط للرعاية المستقبلية.

يتسم مرض باركنسون بتأثيراته البيولوجية على نظام الحركة والأعصاب. يعتبر تلف خلايا الدماغ العصبية، وخاصة في منطقة تُسمى "المادة السوداء"، هي السمة الرئيسية للمرض. هذا التلف يؤدي إلى نقص في إفراز الدوبامين، وهو الناقل العصبي الذي يلعب دوراً حاسماً في التحكم بالحركة (Porter & Gupta, 2023).

تتسبب أعراض مرض باركنسون في تحديات كبيرة للأفراد المصابين، حيث تشمل الرعاش، والذي يجعل التحكم في الحركة الدقيقة أمراً صعباً. بالإضافة إلى ذلك، يمكن أن يتسبب التصلب العضلي وصعوبات التنسيق في إعاقة القدرة على القيام بالأنشطة اليومية بكفاءة.

تتقدم أعراض مرض باركنسون تدريجياً، وتتفاقم مع مرور الوقت. يبدأ المصابون بتجربة أعراض خفيفة، مثل الرعاش الطفيف، وتتطور الأمور تدريجياً إلى مستويات أكثر تعقيداً. الفهم الأعمق لطبيعة المرض يساعد في تحديد الطرق المثلى للتدخل وتحسين نوعية حياة المصابين. على الرغم من أن مرض باركنسون لا يمكن شفاؤه، إلا أن هناك العديد من العلاجات المتاحة التي يمكن أن تساعد في التخفيف من الأعراض.

كما وتلعب العوامل الوراثية دوراً هاماً في ظهور مرض باركنسون. يُعتبر وجود تاريخ عائلي للمرض عاملاً رئيسياً يزيد من احتمالية الإصابة. تشير العديد من الأبحاث العلمية إلى وجود جينات محددة قد تكون مرتبطة بتفاقم المرض (Valente et al., 2016). بجانب العوامل الوراثية، هناك عوامل بيئية يمكن أن تلعب دوراً في تطور مرض باركنسون. يشمل ذلك التعرض للسموم البيئية والمواد الكيميائية في بيئة العمل. يُثبت البحث المستمر أهمية فهم تفاعل هذه العوامل مع الجوانب الوراثية لفهم أفضل للمرض (Valente et al., 2016).

فيما يتعلق بتأثير مرض باركنسون على جهاز الحركة، يتميز هذا الاضطراب العصبي بتأثيره الكبير على نظام الحركة في الجسم. يتسبب انخفاض نشاط مادة الدوبامين في الدماغ، والتي تلعب دوراً رئيسياً في تنظيم

الحركة، في ظهور أعراض تشمل الرعشة، والتي تكون غالباً موجودة في الأطراف العليا، والتي تزداد شدتها مع تقدم المرض. بالإضافة إلى ذلك، تظهر الصعوبة في التحكم بالحركات التلقائية مثل المشي، وقد يتطور التجمد أو التأخر في الحركة، مما يؤدي إلى تأثير بارز على قدرة المصاب بالمرض على أداء الأنشطة اليومية بكفاءة. تُسلط هذه الفقرة الضوء على تأثيرات مرض باركنسون البارزة على الحركة والتحديات التي يواجهها المرضى في تنفيذ حركاتهم بشكل طبيعي، مما يمهد الطريق لفهم أوسع للنواحي الوظيفية والحياتية المتأثرة (Sveinbjornsdottir, 2016).

مرض باركنسون هو اضطراب عصبي تقدمي يؤثر على جهاز الحركة. يتسبب في تدهور التحكم في الحركة ويظهر بشكل رئيسي عند الأفراد الذين تجاوزوا سن الـ 50. من بعض الأعراض والمظاهر المميزة لمرض باركنسون:

✚ رعشة الراحة: تعد الرعشة واحدة من الأعراض البارزة للمرض، حيث يلاحظ المصابين بها رعشة في الأطراف العليا أثناء الراحة. هذه الرعشة تختفي أو تقل عندما يقوم الفرد بالحركة.

✚ صعوبة في التحكم الحركي: يواجه المصابون بمرض باركنسون صعوبة في التحكم الحركي، مما يؤدي إلى تقليل حجم الحركات الطبيعية. الحركات تصبح بطيئة ومحدودة، ويكون من الصعب تنسيق الحركات المعقدة.

✚ تصلب العضلات: يشعر المرضى بتصلب في العضلات، وخاصة في الأطراف. هذا التصلب يجعل الحركة أكثر صعوبة ويساهم في الشعور بالإجهاد والتعب.

✚ فقدان التوازن: تتسبب أعراض مرض باركنسون في فقدان التوازن، مما يزيد من احتمالية الوقوع. المشي يصبح أكثر صعوبة، ويمكن أن يؤدي إلى حوادث سقوط شائعة.

✚ التجمّد أثناء المشي: تظهر ظاهرة تجمّد المريض أثناء المشي، حيث يفقد القدرة على المضي قدماً.

هذه التجمّد قد يستمر لثوانٍ معدودة ويؤثر على قدرة المصاب على التحرك بشكل طبيعي.

✚ تغيرات في التعبير الوجهي: يعاني الأفراد المصابون بالمرض من تغيرات في التعبير الوجهي، حيث

يظهرون عادةً بملامح وجهية ثابتة وقليلة التغير.

✚ صعوبة في التنفس والبلع: تؤثر أعراض المرض على الوظائف الحيوية الأخرى مثل التنفس والبلع،

مما يمكن أن يتسبب في مشاكل صحية إضافية.

✚ اضطرابات النوم: يعاني مرضى باركنسون من اضطرابات في النوم، مثل الأرق والإستيقاظ المتكرر

خلال الليل (Sveinbjornsdottir, 2016).

تتفاوت حدة هذه الأعراض بين المصابين بالمرض، ويمكن أن تتطور مع مرور الوقت. يتطلب التشخيص

الدقيق والعلاج الفعال التعاطف والفهم الواسعين لتأثيرات هذا المرض الشديد.

تحمل الأصوات التي يصدرها الأفراد المصابون بمرض باركنسون تغيرات خاصة بهم، وهي تغيرات صوتية

قد تكون غير ملحوظة بالنسبة للأذن البشرية العادية، ولكن تظهر بوضوح عبر التحليل الصوتي المتقدم.

يمكن أن تشمل هذه التغيرات تقليل في تردد الصوت الأساسي، وتشوهات في نسبة الهمس إلى النبرة، مما

يُميز الأفراد المصابين (Mašić et al., 2018). لاستخدام الصوت في هذا التشخيص عدة مزايا منها:

✚ كشف مبكر: تقدم التحليلات الصوتية الدقيقة إمكانية الكشف المبكر لعلامات مرض باركنسون،

حتى قبل ظهور أعراض بدنية واضحة. هذا يتيح للأطباء التدخل المبكر وتوفير العلاج الملائم في

وقت مبكر، مما يحسن من نتائج العلاج.

✚ فحص غير غازي: تعتبر تقنيات التحليل الصوتي فحوصات غير غازية، مما يسهل على المرضى تنفيذها بشكل دوري دون الحاجة إلى إجراءات طبية معقدة أو مؤلمة. يُعزز هذا النوع من الفحوصات التشخيص المستدام والمريح.

✚ التكلفة المنخفضة: بالمقارنة مع بعض التقنيات الطبية الأخرى، يمكن أن يكون تحليل الصوت تكلفة منخفضة نسبياً، مما يجعله خياراً ملائماً للفحوصات المتكررة ولتوفير الرعاية الصحية بتكلفة معقولة. دقة التشخيص: توفر تقنيات تحليل الصوت دقة عالية في التشخيص، حيث يمكنها تمييز التغيرات الصوتية الدقيقة المرتبطة بمرض باركنسون. هذا يعزز القدرة على تحديد الحالات المصابة بدقة وتوجيه العلاج بفعالية (Mašić et al., 2018).

يعتبر تحليل الصوت أداة فعالة وغير غازية في تشخيص مرض باركنسون، حيث يجمع بين الكفاءة والتكلفة المنخفضة والدقة في التحليل.

تعتمد الأساليب التقليدية للكشف عن مرض باركنسون على الإختبارات السريرية والتقييمات الطبية التي يقوم بها الأطباء المختصون. تشمل هذه الإختبارات قياسات حركة الجسم، واختبارات الثبات والتوازن، وتقييمات الوظائف الحركية الأخرى. يعتمد هذا النوع من الكشف على ملاحظات وتقييمات الأطباء بناءً على تفاعل المريض في الزمن الحالي (Sveinbjornsdottir, 2016). ولاستخدام الاساليب التقليدية عدة تحديات، منها:

✚ تبعات زمنية: يمكن أن تتطلب الإختبارات التقليدية والتقييمات الطبية وقتاً طويلاً للتنفيذ، وقد تكون مكلفة أحياناً، مما يؤدي إلى تأخير في التشخيص والبدء في العلاج.

الدقة المحدودة: تعتمد فعالية الكشف عن مرض باركنسون باستخدام الأساليب التقليدية على خبرة الطبيب وقدرته على تفسير التفاعلات والأعراض التي يلاحظها. قد تكون هناك قضايا متعلقة بالدقة في حالات التشخيص المبكرة أو الخفيفة.

الإعتماد على التفاعلات الحالية: تتطلب الأساليب التقليدية وجود التفاعلات الحالية للمريض، مما قد يجعلها غير فعالة في تشخيص الحالات التي قد لا تظهر أعراضاً واضحة في اللحظة الراهنة (Mou et al., 2020).

تظهر الأساليب التقليدية للكشف عن مرض باركنسون بعض التحديات فيما يتعلق بالتكلفة والزمن والدقة. من هنا، تبرز أهمية إستكمال هذه الأساليب بوسائل تحليل الصوت المتقدمة لضمان تشخيص دقيق وفعال.

1.5 استخدام أدوات الذكاء الاصطناعي في مرض باركنسون

تعد تقنيات التعلم الآلي ومعالجة اللغات الطبيعية جزءاً أساسياً من تطورات مجال الذكاء الاصطناعي والتحليل البياني. في سياق مرض باركنسون، يُستخدم التعلم الآلي لتحليل نماذج معقدة من البيانات الصوتية المستخرجة من تسجيلات الأفراد. هذا يشمل تحديد التغيرات الدقيقة في التردد الصوتي ونغمة الصوت، التي يمكن أن تكون مرتبطة بتطورات المرض. وتلعب معالجة اللغات الطبيعية دوراً في فهم السياق اللغوي للتسجيلات الصوتية، مما يمكن من تحليل النطق والتعبير بشكل أفضل. هذا يعزز قدرة النظم على تفسير التغيرات اللفظية التي قد تظهر في حالات الباركنسون (Govindu & Palwe, 2023).

تقدم التقنيات الحديثة إمكانيات فريدة للكشف المبكر عن مرض باركنسون، وذلك عبر تحليل دقيق وشامل للبيانات الصوتية. في مراحل مبكرة، قد لا تكون الأعراض واضحة، ولكن باستخدام التكنولوجيا، يمكن إكتشاف تغيرات طفيفة في النمط الصوتي قبل ظهور أي علامات ظاهرة. كما وتسهم هذه التقنيات في توجيه

الرعاية الطبية بشكل أفضل، حيث يتيح التشخيص المبكر فرصاً أكبر للتدخل العلاجي، مما يسهم في تحسين نوعية حياة المرضى. تكنولوجيا التحليل الصوتي تمثل بوابة لمستقبل واعد في تشخيص ومعالجة أمراض الجهاز العصبي مثل مرض باركنسون. حيث يمكن لتقنيات معالجة الصوت الحديثة أن تستخرج المميزات المختلفة والعميقة من التسجيلات الصوتية حيث هذه المميزات هي التي تسهم في معرفة مزايا هذا التسجيل وخواصها والتعرف على بعض أمراض أو سلوكيات صاحب التسجيل من خلالها. وتتجلى أهمية ذلك في عدّة أمور:

✚ فعالية بيانات التسجيل في تحليل التغيرات الصوتية

بيانات التسجيل الصوتي تمثل مصدراً غنياً للمعلومات، حيث تحتوي على تفاصيل دقيقة حول النمط الصوتي للأفراد. يُستخدم تحليل هذه البيانات لفحص التغيرات الدقيقة في التردد والنبرة والنغمة، مما يساعد في إكتشاف أي تغيرات صوتية قد ترتبط بتقدم مرض باركنسون. تقف هذه البيانات كشاهد حي على النمط الصوتي للأفراد، مما يمنحنا إمكانية فحص التغيرات الصوتية بدقة فائقة (Mašić et al., 2018). تكنولوجيا معالجة الصوت وتعلم الآلة تمكن من إستخراج نقاط بيانية دقيقة من بيانات التسجيل، مما يُمكن الأطباء والباحثين من فحص النمط الصوتي بشكل شامل وفعال. حيث يُفترض أن مرض باركنسون يترافق مع تغيرات دقيقة في النبرة والتردد الصوتي، والتي قد لا تلاحظها الأذن البشرية بشكل واضح. هنا يبرز دور تحليل بيانات التسجيل الصوتي، حيث يُمكن لتقنيات معالجة الصوت والتعلم الآلي إستخراج هذه التغيرات الصوتية الدقيقة وتحليلها بشكل شامل.

✚ تمييز بيانات المصابين وغير المصابين بشكل فعال

في سعينا لتحسين تشخيص مرض باركنسون، تكمن أهمية فحص بيانات التسجيل الصوتي في قدرتها على التمييز بين الأفراد المصابين وغير المصابين بشكل دقيق. بفضل تطور التقنيات، يمكن

إستخدام نماذج التعلم الآلي لتحديد الأنماط الصوتية الفريدة التي تميز مرضى باركنسون عن غيرهم. حيث ما يميز بيانات التسجيل الصوتي هو قدرتها على التمييز بين الأفراد المصابين بمرض باركنسون والأفراد الأصحاء بشكل دقيق. بفضل تقدم تقنيات التعلم الآلي، يمكن أن تعتمد النماذج على الأنماط الصوتية الفريدة التي تعكس التغيرات الناتجة عن المرض. يتيح هذا التمييز الدقيق تحديد المرض في مراحله المبكرة، مما يفتح أفقاً جديدة لتحسين فرص العلاج المبكر والإدارة الفعّالة للحالة. إن استخدام بيانات التسجيل الصوتي كأداة للتشخيص ليس فقط يُسهم في فهمنا لمرض باركنسون بل يمثل أيضاً اتجاهاً نحو مستقبل طبي يعتمد على تكنولوجيا الصوت.

لماذا استخدمنا بيانات مستخرجة من تسجيلات صوتية؟

استخدام بيانات تسجيلات صوتية في تشخيص مرض باركنسون يعود إلى الحقيقة الطبية التي تشير إلى أن هناك تغيرات في النبذة والتوتر الصوتي يمكن أن تكون مرتبطة بأمراض معينة، بما في ذلك مرض باركنسون. في حالة مرض باركنسون، قد تحدث تغيرات في النبذة والقوة الصوتية نتيجة للتأثيرات على الجهاز الحركي والتحكم العصبي (Govindu & Palwe, 2023).

تحليل تسجيلات الصوت يمكن أن يساعد في اكتشاف تلك التغيرات الدقيقة وتقديم مؤشرات على وجود مشكلات صوتية مرتبطة بالمرض. البيانات الصوتية يمكن أن تحتوي على معلومات ثرية تفصل بين الخصائص الصوتية المختلفة التي يمكن أن تكون مفيدة في تشخيص حالات معينة، بما في ذلك مرض باركنسون.

1.6 التعلم الآلي machine learning

التعلم الآلي هو أحد تطبيقات الذكاء الاصطناعي (AI) الذي يوفر للأنظمة القدرة على التعلم والتحسين تلقائياً من التجربة دون أن تتم برمجتها بشكل صريح.

يركّز التعلّم الآلي على تطوير برامج الكمبيوتر التي يمكنها الوصول إلى البيانات واستخدامها للتعلم بأنفسهم. تبدأ عملية التعلّم بالملاحظات أو البيانات، مثل الأمثلة أو الخبرة المباشرة أو التعليمات، من أجل البحث عن أنماط في البيانات وإتخاذ قرارات أفضل في المستقبل بناءً على الأمثلة التي نقدمها.

الهدف الأساسي هو السماح لأجهزة الكمبيوتر بالتعلّم تلقائياً دون تدخل بشري أو مساعدة وتعديل الإجراءات وفقاً لذلك. ولكن باستخدام الخوارزميات الكلاسيكية للتعلم الآلي، يعتبر النص بمثابة سلسلة من الكلمات الرئيسية؛ بدلاً من ذلك، النهج القائم على التحليل الدلالي يُحاكي قدرة الإنسان على فهم معنى النص.

و"الخوارزمية" في التعلّم الآلي هي إجراء يتم تشغيله على البيانات لإنشاء "نموذج" للتعلّم الآلي. تؤدي خوارزميات التعلم الآلي "التعرف على الأنماط". الخوارزميات "تتعلم" من البيانات، أو "ملائمة" لمجموعة البيانات. هناك العديد من خوارزميات التعلّم الآلي. على سبيل المثال، لدينا خوارزميات للتصنيف، مثل k -أقرب الجيران. لدينا خوارزميات للانحدار، مثل الانحدار الخطي، ولدينا خوارزميات للتجميع، مثل k -mean.

و"النموذج" (model) في التعلم الآلي هو ناتج خوارزمية التعلم الآلي التي تعمل على البيانات. ويمثّل النموذج ما تعلمته خوارزمية التعلم الآلي. النموذج هو "الشيء" الذي يتم حفظه بعد تشغيل خوارزمية التعلم الآلي على بيانات التدريب ويمثّل القواعد والأرقام وأي هياكل بيانات أخرى خاصة بالخوارزمية المطلوبة لعمل تنبؤات (Vakili et al., n.d). حيث قد توضح بعض الأمثلة ذلك:

- ينتج عن خوارزمية الانحدار الخطي نموذج يتكون من متجه للمعاملات بقيم محددة.
 - ينتج عن خوارزمية شجرة القرار نموذجًا يتألف من شجرة من عبارات if-then بقيم محددة.
 - تؤدي خوارزميات الشبكة العصبية / الانتشار العكسي / والانحدار معًا إلى نموذج يتكون من بنية رسم بياني مع متجهات أو مصفوفات من الأوزان بقيم محددة.
- يعدّ نموذج التعلم الآلي أكثر صعوبة بالنسبة للمبتدئين لأنه لا يوجد تشابه واضح مع الخوارزميات الأخرى في علوم الكمبيوتر. فعلى سبيل المثال، ناتج القائمة التي تم فرزها لخوارزمية الفرز ليس نموذجًا بالفعل. أفضل تشبيه هو التفكير في نموذج التعلم الآلي على أنه "برنامج". ويتكون "برنامج" نموذج التعلم الآلي من كل من البيانات وإجراء لاستخدام البيانات لعمل تنبؤ. على سبيل المثال، ضع في إعتبارك خوارزمية الانحدار الخطي والنموذج الناتج. حيث يتكون النموذج من متجه من المعاملات (البيانات) التي يتم ضربها وتلخيصها بصف من البيانات الجديدة المأخوذة كمدخلات من أجل إجراء التنبؤ.
- نقوم بحفظ البيانات الخاصة بنموذج التعلم الآلي لاستخدامه لاحقًا. غالباً ما نستخدم إجراء التنبؤ لنموذج التعلم الآلي الذي توفره مكتبة التعلم الآلي.
- في بعض الأحيان قد نقوم بتنفيذ إجراء التنبؤ بأنفسنا كجزء من تطبيقنا. غالباً ما يكون هذا أمراً بسيطاً نظراً لأن معظم إجراءات التنبؤ بسيطة للغاية.

1.7 خوارزميات التعلم الآلي شائعة الاستخدام

فيما يلي قائمة بخوارزميات التعلم الآلي الشائعة الاستخدام. يمكن تطبيق هذه الخوارزميات على أي مشكلة بيانات تقريباً:

الإنحدار الخطي (Linear Regression)

الإنحدار اللوجستي (Logistic Regression)

شجرة القرار (Decision Tree)

SVM

ساذج بايز (Naïve Bayes)

kNN (K-Nearest neighbors)

K-الوسائل (k-medoids)

الغابة عشوائية (Random Forest)

خوارزميات تخفيض الأبعاد

خوارزميات تعزيز التدرج

1.7.1 الإنحدار اللوجستي logistic regression

في الإحصائيات، يتم استخدام النموذج اللوجستي (أو نموذج اللوجستيات) لنمذجة احتمالية فئة معينة أو حدث موجود مثل النجاح / الفشل، الفوز / الخسارة، على قيد الحياة / ميت، أو سليم / مريض. يمكن توسيع هذا ليشمل عدة فئات من الأحداث مثل تحديد ما إذا كانت الصورة تحتوي على قطة أو كلب أو أسد، إلخ. سيتم تعيين احتمال بين 0 و 1 لكل كائن في الصورة، بمجموع واحد.

الإنحدار اللوجستي هو نموذج إحصائي يستخدم في شكله الأساسي وظيفة لوجستية لنمذجة متغير ثنائي تابع، على الرغم من وجود العديد من الإمتدادات الأكثر تعقيداً. في تحليل الإنحدار، يقوم الإنحدار اللوجستي بتقدير معلمات النموذج اللوجستي (شكل من أشكال الإنحدار الثنائي). رياضياً، يحتوي النموذج اللوجستي الثنائي على متغير تابع بقيمتين محتملتين، مثل النجاح / الفشل الذي يتم تمثيله بواسطة متغير مؤشر، حيث يتم تسمية القيمتين "0" و "1". في النموذج اللوجستي، فإن الاحتمالات اللوغاريتمية (لوغاريتم الاحتمالات) للقيمة المسماة "1" هي مزيج خطي من واحد أو أكثر من المتغيرات المستقلة (تنبؤات)، يمكن أن تكون كل من المتغيرات المستقلة متغيراً ثنائياً (فئتان، مشفرة بواسطة متغير مؤشر) أو متغير مستمر (أي قيمة حقيقية).

يمكن أن يختلف الاحتمال المقابل للقيمة المسماة "1" بين 0 (بالتأكيد القيمة "0") و "1" (بالتأكيد القيمة "1")، ومن ثم وضع العلامات؛ الوظيفة التي تحول احتمالات السجل إلى احتمالية هي الوظيفة اللوجستية، ومن هنا جاء الاسم. تسمى وحدة القياس لمقياس الاحتمالات اللوغاريتمية logit، من الوحدة اللوجستية، ومن هنا جاءت الأسماء البديلة.

كما ويمكن أيضاً استخدام نماذج مماثلة ذات وظيفة سينية مختلفة بدلاً من الوظيفة اللوجستية، مثل نموذج الاحتمال؛ السمة المميزة للنموذج اللوجستي هي أن زيادة أحد المتغيرات المستقلة تؤدي بشكل مضاعف إلى زيادة احتمالات النتيجة المعطاة عند معدل ثابت، مع كل متغير مستقل له معلمة خاصة به؛ بالنسبة للمتغير الثنائي التابع، فإن هذا يعمم نسبة الأرجحية.

في نموذج الإنحدار اللوجستي الثنائي، يحتوي المتغير التابع على مستويين (فئتين). يتم نمذجة المخرجات التي تحتوي على أكثر من قيمتين من خلال الإنحدار اللوجستي متعدد الحدود، وإذا تم ترتيب الفئات

المتعددة، عن طريق الإنحدار اللوجستي الترتيبي (على سبيل المثال نموذج الإحتمالات النسبية الترتيبي اللوجستي).

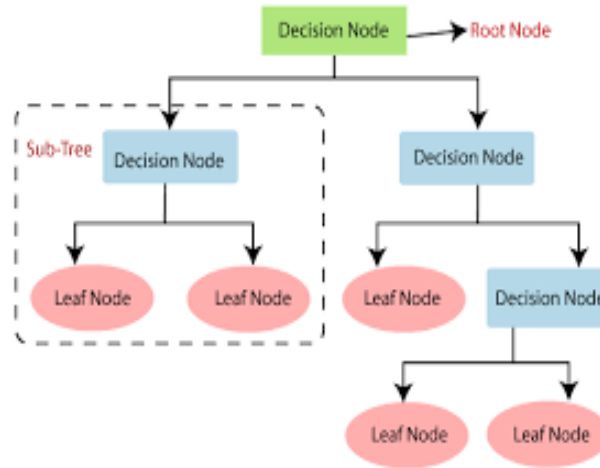
نموذج الإنحدار اللوجستي نفسه يصوغ ببساطة إحتمالية المخرجات من حيث المدخلات ولا يقوم بتصنيف إحصائي (ليس مصنفاً)، على الرغم من أنه يمكن استخدامه لإنشاء مصنف، على سبيل المثال عن طريق إختيار قيمة القطع وتصنيف المدخلات باستخدام إحتمالية أكبر من الحد كفاءة واحدة، أقل من القطع مثل الأخرى؛ هذه طريقة شائعة لعمل مصنف ثنائي (Sperandei, 2014).

1.7.2 شجرة القرار

شجرة القرار هي تقنية تعلم خاضعة للإشراف يمكن استخدامها لكل من مشاكل التصنيف والإنحدار، ولكنها في الغالب مفضلة لحل مشاكل التصنيف. إنه مصنف منظم على شكل شجرة، حيث تمثل العقد الداخلية ميزات مجموعة البيانات، وتمثل الفروع قواعد القرار وتمثل كل عقدة ورقية النتيجة.

في شجرة القرار، توجد عقدتان، وهما عقدة القرار والعقدة الورقية. تُستخدم عقد القرار لاتخاذ أي قرار ولها فروع متعددة، في حين أن العقد الورقية هي نتاج تلك القرارات ولا تحتوي على أي فروع أخرى. ويتم تنفيذ القرارات أو الإختبار على أساس ميزات مجموعة البيانات المحددة. إنها تمثيل رسومي للحصول على جميع الحلول الممكنة لمشكلة / قرار بناءً على شروط معينة (Guia et al., 2019).

يطلق عليها شجرة القرار لأنها تشبه الشجرة، فهي تبدأ بالعقدة الجذرية، والتي تتوسع في المزيد من الفروع وتبني بنية شبيهة بالشجرة. من أجل بناء شجرة، نستخدم خوارزمية CART، والتي تعني التصنيف وخوارزمية شجرة الإنحدار. تطرح شجرة القرار سؤالاً ببساطة واستناداً إلى الإجابة (نعم / لا)، فإنها تقسم الشجرة إلى أشجار فرعية. ويوضح الرسم البياني 1 أدناه الهيكل العام لشجرة القرار.



الشكل 1 الهيكل العام لشجرة القرار DT (Guia et al., 2019)

1.7.3 خوارزمية SVM

تعد Support Vector Machine أو SVM واحدة من أكثر خوارزميات التعلم الخاضع للإشراف شيوعاً، والتي تُستخدم في التصنيف بالإضافة إلى مشاكل الانحدار. ومع ذلك، في المقام الأول، يتم استخدامه لمشاكل التصنيف في التعلم الآلي.

الهدف من خوارزمية SVM هو إنشاء أفضل خط أو حدود قرار يمكنها فصل الفضاء ذي البعد n إلى فئات بحيث يمكننا بسهولة وضع نقطة البيانات الجديدة في الفئة الصحيحة في المستقبل. يُطلق على حدود القرار الأفضل اسم المستوى الفائق.

يختار SVM النقاط / المتجهات القصوى التي تساعد في إنشاء المستوى الفائق. تسمى هذه الحالات القصوى متجهات الدعم، وبالتالي تسمى الخوارزمية بآلة متجه الدعم.

تأمل المثال التالي. لنفترض أننا رأينا قطاً غريباً يحتوي أيضاً على بعض ميزات الكلاب، لذلك إذا أردنا نموذجاً يمكنه تحديد ما إذا كان قطة أو كلباً بدقة، فيمكن إنشاء مثل هذا النموذج باستخدام خوارزمية

SVM. سنقوم أولاً بتدريب نموذجنا بالكثير من صور القطط والكلاب حتى يتمكن من التعرف على الميزات المختلفة للقطط والكلاب، ثم نقوم باختباره مع هذا المخلوق الغريب. لذا، نظراً لأن متجه الدعم ينشئ حدوداً للقرار بين هاتين المعطيات (القط والكلب) ويختار الحالات القصوى (متجهات الدعم)، فسوف يرى الحالة القصوى للقطط والكلب. على أساس ناقلات الدعم، سوف يصنفها على أنها قطة (Guia et al., 2019).

1.7.4 خوارزمية Naïve Bayes

خوارزمية Naïve Bayes هي خوارزمية تعلم خاضعة للإشراف، والتي تستند إلى نظرية بايز وتستخدم لحل مشاكل التصنيف. يتم استخدامه بشكل أساسي في تصنيف النص الذي يتضمن مجموعة بيانات تدريب عالية الأبعاد.

يعد تصنيف Naïve Bayes أحد خوارزميات التصنيف البسيطة والأكثر فعالية والتي تساعد في بناء نموذج تعلم آلي سريعة يمكنها إجراء تنبؤات سريعة. إنه مصنف احتمالي، مما يعني أنه يتنبأ على أساس احتمال وجود كائن.

بعض الأمثلة الشائعة لخوارزمية Naïve Bayes هي ترشيح البريد العشوائي والتحليل العاطفي وتصنيف المقالات.

تتكون خوارزمية Naïve Bayes من كلمتين Naïve و Bayes، والتي يمكن وصفها على النحو التالي:

🌈 ساذج: يطلق عليه Naïve لأنه يفترض أن حدوث سمة معينة مستقلة عن حدوث ميزات أخرى.

على سبيل المثال، إذا تم تحديد الفاكهة على أساس اللون والشكل والمذاق، فإن الفاكهة الحمراء والكروية والحلوة يتم التعرف عليها على أنها تفاحة. ومن ثم تساهم كل ميزة على حدة في تحديد أنها تفاحة دون الاعتماد على بعضها البعض.

بايز: يطلق عليه Bayes لأنه يعتمد على مبدأ نظرية بايز. تُعرف نظرية بايز أيضاً باسم قانون

بايز، والذي يستخدم لتحديد احتمالية الفرضية بمعرفة مسبقة. يعتمد على الإحتمال الشرطي (Guia

et al., 2019). تُعطى صيغة نظرية بايز على النحو التالي:

$$p(A_i|B) = \frac{p(A_i).p(B|A_i)}{p(A_1).p(B|A_1) + p(A_2).p(B|A_2) + \dots + p(A_n).p(B|A_n)}$$

1.7.5 خوارزمية KNN

K-Nearest Neighbor هي واحدة من أبسط خوارزميات التعلم الآلي القائمة على تقنية التعلم الخاضع

للإشراف. تفترض خوارزمية KNN التشابه بين الحالة / البيانات الجديدة والحالات المتاحة وتضع الحالة

الجديدة في الفئة الأكثر تشابهاً مع الفئات المتاحة.

تقوم خوارزمية KNN بتخزين جميع البيانات المتاحة وتصنيف نقطة بيانات جديدة بناءً على التشابه. هذا

يعني أنه عند ظهور بيانات جديدة يمكن تصنيفها بسهولة إلى فئة مجموعة جيدة باستخدام خوارزمية KNN.

يمكن استخدام هذه الخوارزمية للانحدار وكذلك للتصنيف ولكن في الغالب يتم استخدامه لمشاكل التصنيف.

إنها خوارزمية غير معلمية، مما يعني أنها لا تضع أي افتراضات بشأن البيانات الأساسية. يُطلق عليها

أيضاً خوارزمية المتعلم الكسول لأنها لا تتعلم من مجموعة التدريب على الفور بدلاً من ذلك تقوم بتخزين

مجموعة البيانات وفي وقت التصنيف، ثم تقوم بتنفيذ إجراء على مجموعة البيانات (Putra et al.,

2022).

تقوم خوارزمية KNN في مرحلة التدريب فقط بتخزين مجموعة البيانات وعندما تحصل على بيانات جديدة،

فإنها تصنف تلك البيانات في فئة تشبه إلى حد كبير البيانات الجديدة. لنفترض على سبيل المثال، أن لدينا

صورة لمخلوق يشبه قطة وكلب، لكننا نريد أن نعرف إما أنه قطة أو كلب. لذلك من أجل هذا التعريف،

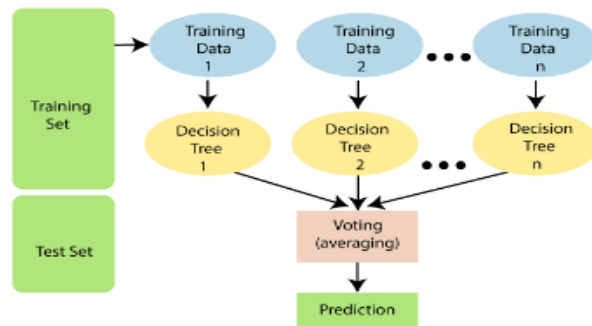
يمكننا استخدام خوارزمية KNN، لأنها تعمل على مقياس التشابه. سيجد نموذج KNN الخاص بنا الميزات المماثلة لمجموعة البيانات الجديدة لصور القطط والكلاب واستناداً إلى الميزات الأكثر تشابهاً، سيضعها في فئة القطط أو الكلاب (Putra et al., 2022).

1.7.6 الغابة العشوائية

Random Forest هي خوارزمية تعلم آلي شائعة تنتمي إلى تقنية التعلم الخاضع للإشراف. يمكن استخدامها لكل من مشاكل التصنيف والانحدار في ML. يعتمد على مفهوم التعلم الجماعي، وهو عملية الجمع بين عدة مصنفات لحل مشكلة معقدة وتحسين أداء النموذج.

حسب الاسم المقترح، Random Forest هو مصنف يحتوي على عدد من أشجار القرار في مجموعات فرعية مختلفة من مجموعة البيانات المحددة ويأخذ المتوسط لتحسين الدقة التنبؤية لمجموعة البيانات هذه. بدلاً من الاعتماد على شجرة قرار واحدة، تأخذ الغابة العشوائية التنبؤ من كل شجرة وتعتمد على أغلبية الأصوات للتنبؤات، وتتوقع الناتج النهائي (Breiman, 2001).

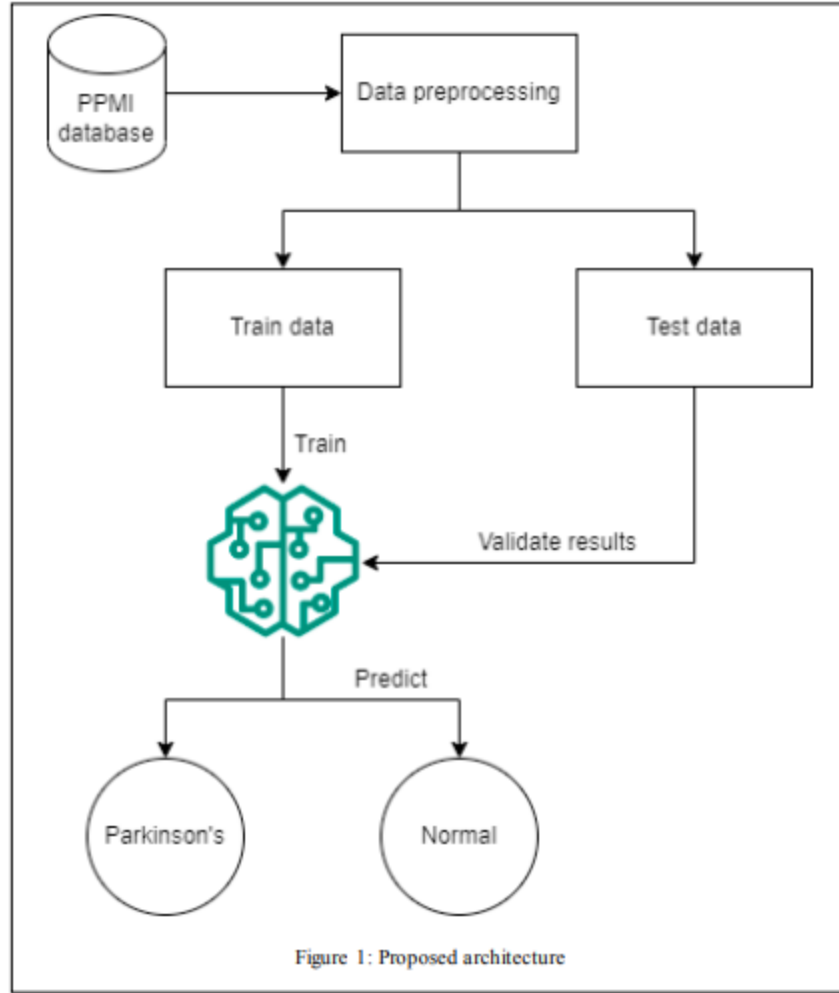
بدلاً من الاعتماد على شجرة قرار واحدة، تأخذ الغابة العشوائية التنبؤ من كل شجرة وتعتمد على أغلبية الأصوات للتنبؤات، وتتوقع الناتج النهائي. يوضح الشكل 2 أدناه عمل خوارزمية Random Forest.



الشكل 2 عمل خوارزمية Random Forest (Breiman, 2001).

2 الفصل الثاني: الدراسة المرجعية

قام Govindu وآخرون بدراسة (Govindu & Palwe, 2023) بعنوان Early detection of Parkinson's disease using machine learning حيث قاموا باكتشاف مرض باركنسون وتصنيف المرضى إلى أصحاء ومصابين باستخدام بيانات تحتوي على معلومات حول الإرتعاش والوميض و MDVP لنطق حروف العلة. قاموا بمعالجة البيانات مسبقاً قبل التدريب وتحليلها وتصورها من أجل فهم شامل للسمات. ثم قاموا بتدريب أربعة نماذج - الإنحدار اللوجستي logistic regression، و SVM، و Random Forest Regressor، و K أقرب الجيران - على 75% من البيانات. تم تدريب النماذج على تصنيف البيانات الصوتية المعطاة إلى PD أو سليمة، بناءً على الاختلافات في التردد. يتم إختبار النماذج على 25% من البيانات ويتم تقييمها على أساس الحساسية sensitivity والدقة accuracy والدقة precision ومصفوفة الإرتباك confusion matrix ودرجة ROC-AUC. الشكل 3 يوضح المخطط الصندوقي لمراحل العمل التي قاموا بها.



الشكل 3 المخطط الصندوقي لمراحل عمل الدراسة المرجعية (Govindu & Palwe, 2023)

قاموا باكتشاف مرض باركنسون باستخدام ثلاثة منهجيات للعمل بعد ما قاموا بموازنة مجموعة البيانات قبل البدء حيث أن البيانات غير متوازنة كانت في البداية وهذه المنهجيات كالتالي:

2.1 المنهجية الأولى

في المنهجية الأولى تم التدريب على 22 مميزة كالتالي:

- ✚ جمع البيانات الصوتية MDVP من قواعد بيانات UCI و PPPMI
- ✚ إجراء تحليل البيانات للكشف عن الانحراف skew وعدم التوازن وتوزيع المتغيرات في البيانات
- ✚ قياس البيانات إلى نطاق مشترك باستخدام المقياس القياسي standard scaler
- ✚ تقسيم مجموعة البيانات إلى مجموعات إختبار وتدريب، حيث تمثل بيانات التدريب 75% من الإجمالي

✚ تدريب نماذج SVM والانحدار اللوجستي والغابات العشوائية ونماذج KNN (Govindu & Palwe, 2023).

2.2 المنهجية الثانية

يتم تطبيق تحليل المكونات الرئيسية (PCA) لتحديد 5 مميزات رئيسية كالتالي:

- ✚ جمع البيانات الصوتية MDVP من قواعد بيانات UCI و PPPMI
- ✚ إجراء تحليل البيانات للكشف عن الانحراف skew وعدم التوازن وتوزيع المتغيرات في البيانات
- ✚ قياس البيانات إلى نطاق مشترك باستخدام المقياس القياسي standard scaler
- ✚ تحديد التباين في كل عمود من البيانات وتطبيق تحليل المكونات الرئيسية (PCA) لتحديد 5 من الميزات الأكثر صلة بنموذج التدريب، من بين 22 سمة.

✚ تقسيم مجموعة البيانات إلى مجموعات إختبار وتدريب، حيث تمثل بيانات التدريب 75% من الإجمالي

✚ إعادة تدريب نماذج SVM والانحدار اللوجستي والغابات العشوائية ونماذج KNN.

✚ مقارنة نتائج التصنيف باستخدام مصفوفة الارتباك ومنحنى ROC-AUC والدقة. (Govindu & Palwe, 2023)

2.3 المنهجية الثالثة

قاموا بإزالة الخل من مجموعة البيانات كالتالي:

✚ جمع البيانات الصوتية MDVP من قواعد بيانات UCI و PPPMI

✚ إجراء تحليل البيانات للكشف عن الانحراف skew وعدم التوازن وتوزيع المتغيرات في البيانات

✚ إكتشفوا أن مجموعة البيانات غير متوازنة، حيث يوجد 109 سجلات لأشخاص مصابين و 40 سجلاً لأشخاص أصحاء، فقاموا بموازنتها.

✚ قياس البيانات إلى نطاق مشترك باستخدام المقياس القياسي standard scaler

✚ تقسيم مجموعة البيانات إلى مجموعات إختبار وتدريب، حيث تمثل بيانات التدريب 75% من الإجمالي

✚ إعادة تدريب نماذج SVM والانحدار اللوجستي والغابات العشوائية ونماذج KNN.

✚ مقارنة نتائج التصنيف باستخدام مصفوفة الارتباك ومنحنى ROC-AUC والدقة

لتحديد أفضل نموذج، قاموا بمقارنة نتائج 3 أساليب و 9 نماذج قاموا بتدريبها. المقاييس المختارة هي منحنى ROC-AUC، ومصفوفة الارتباك، والدقة accuracy، والضبط Precision، والإستدعاء Recall، ودرجة F1 (f1-score). يعطي تصنيف مرض باركنسون دقة تبلغ 91.835% وحساسية تبلغ 0.95 لمصنف Random Forest. تعتبر نتائج نموذج Random Forest مثالية، نظراً للأهمية المتساوية الممنوحة لجميع السمات الـ 22 في مجموعة بيانات MDVP. تسلط هذه الورقة الضوء أيضاً على نتائج نموذج SVM الذي يعطي دقة تبلغ 91.836% وحساسية تبلغ 0.94، بعد تطبيق PCA على مجموعة البيانات. يعمل كل من نموذجي SVM و Random Forest جيداً بالنسبة للقيم المتطرفة (Govindu & Palwe, 2023).

يمكن الإستفادة من هذه الدراسة التي تعدّ الخطوة الأولى قبل تنفيذ المشروع حيث تعتبر نظرة عامة على أهم نماذج التعلم الآلي المستخدمة لحل هذه المشكلة مع هذه البيانات، كما تساعدنا على فهم طبيعة البيانات وكيفية إستخراج هذه المميزات وفهمها. أيضاً نستطيع من خلالها فهم مراحل العمل التي سيمرّ بها المشروع بدءاً من قراءة مجموعة البيانات وفهمها حتى مرحلة إختبار النماذج وتقييمها. وبالنظر إلى هذه المنهجيات عن كثب نلاحظ أنها متشابهة في المنحى العام للعمل ومتشابهة في بعض الخطوات ولكن المنهجية الثالثة تعتبر الأفضل لأن المنهجية الأولى والثانية لم يتم فيها موازنة البيانات قبل تدريب النماذج ونحن نعلم أن البيانات يجب أن تكون متوازنة حتى لا يتحيز النموذج إلى صف معين على حساب صف آخر. ولكن من الممكن لنا أن نقوم ببعض التعديلات والإضافات عليها سعياً للتحسين والحصول على نتائج أفضل، ممكن مثلاً استخدام تقنيات أخرى لاختيار بعض المميزات، وتدريب خوارزميات أخرى غير التي تم تدريبها في هذه الدراسة وبعض التعديلات على عملية معالجة البيانات، لأن أي تغيير في أي خطوة من العملية يمكن أن يؤثر على نتيجة العمل.

3 الفصل الثالث: مجموعة البيانات ومعايير التقييم

3.1 مجموعة البيانات dataset

تم استخدام مجموعة البيانات الموجودة على موقع Kaggle، يمكن الوصول إليها من [هنا](#). تتكون مجموعة البيانات هذه من مجموعة من قياسات الصوت الطبية الحيوية لـ 31 شخصاً، 23 منهم يعانون من مرض باركنسون (PD). يمثل كل عمود في الجدول مقياساً صوتياً معيناً، ويتوافق كل صف مع واحد من 195 تسجيلاً صوتياً لهؤلاء الأفراد (عمود "name"). الهدف الرئيسي من البيانات هو التمييز بين الأشخاص الأصحاء والذين يعانون من مرض باركنسون، وفقاً لعمود "status" الذي تم ضبطه على 0 للأصحاء و1 لمرضى باركنسون.

البيانات بتنسيق ASCII CSV. تحتوي صفوف ملف CSV على مثل يتوافق مع تسجيل صوتي واحد. يوجد حوالي ستة تسجيلات لكل مريض، ويتم تحديد اسم المريض في العمود الأول. الجدول 1 يوضح وصف مميزات مجموعة البيانات (Little et al., n.d).

الجدول 1 وصف مميزات مجموعة البيانات (Govindu & Palwe, 2023)

المميزات features	الوصف
Name	يتم تخزين البيانات بتنسيق ASCII CSV حيث يتم حفظ المريض. يتم تخزين الاسم ورقم التسجيل
MDVP: Fo (Hz)	التردد الأساسي لفترة الصوت
MDVP: Fhi (Hz)	الحد الأعلى للتردد الأساسي أو الحد الأقصى لتعديل الصوت
MDVP: Flo (Hz)	الحد الأدنى لتردد الصوت الأساسي للصوت
MDVP: Jitter, Abs, RAP, PPQ, DDP	هذه قياسات متعددة الأبعاد لبرنامج Kay Pentax للصوت MDVP. MDVP هو قياس تقليدي لتردد الإهتزازات في الحبال الصوتية خلال فترة النغمة إلى

	الاهتزازات في بداية دورة صوتية تسمى العلامة النغمية (pitch mark).
Jitter and Shimmer	قياسات الفارق المطلق بين ترددات كل دورة، بعد توحيد المتوسط.
NHR and HNR	قياسات نسبة الإشارة إلى الضوضاء والنسبة الصوتية، التي تشير إلى قوة البيئة في مقاومة التشويش.
D2	تُستخدم بُعد الترابط لتحديد حالة اضطراب النطق باستخدام كائنات ترابطية. إنها خاصية غير خطية وديناميكية.
RPDE	تقوم إنتروبية كثافة فترة التكرار بقياس مدى إنتظام الإشارة أو الإيقاع فيها.
DFA	تقوم تحليل التذبذب المنحرف أو DFA بقياس مدى التشابه الذاتي العشوائي للضوضاء في إشارات الكلام.
PPE	يُستخدم إنتروبيا فترة النغمة لتقييم التغيرات غير الطبيعية في الكلام على مقياس لوغاريتمي.
Spread1, spread2	تحليل نطاق أو مدى التغيرات في الكلام بالنسبة لـ MDVP: Fo (هرتز).
status	الصفر (0) يشير إلى شخص سليم، في حين أن الواحد (1) يشير إلى شخص يعاني من مرض باركنسون (PWP).

المعلومات الموجودة في الجدول السابق استناداً إلى (Little et al., n.d) وهي المقالة العلمية التي توضح مجموعة البيانات التي قمنا باستخدامها. الجدول يوضح تفاصيل كل مميزة من المميزات وتأثيرها على اكتشاف مرض باركنسون، حيث أنّ قيم هذه المميزات ستختلف من صوت شخص لصوت شخص آخر بناءً

على قوة صوته والبيئة المحيطة به ومخارج الحروف والكلام، فالأشخاص المصابين بمرض باركنسون قد يجدون صعوبة في الحفاظ على الإيقاع والتوازن في تقديم الكلمات، وفي إنتاج الكلمات بوضوح، مما يمكن أن يؤدي إلى تشوش في النطق. هذا التشوه يظهر بقيم مختلفة في المميزات الموجودة في مجموعة البيانات وستقوم نماذج التعلم الآلي بملاحظة هذا الفرق بين السجلات التي تصف الأشخاص الأصحاء والسجلات التي تصف الأشخاص المصابين ومن هذه الملاحظات سنقوم باكتشاف الأشخاص المصابين.

3.2 معايير التقييم Evaluation matrices

تتعدد خوارزميات وطرق التقييم والمفاضلة بين خوارزميات التعلم الآلي، ونقصد بتقييم الخوارزمية هو معرفة مدى قوة ومتانة الخوارزمية عندما تتعامل مع بيانات جديدة غير التي تدربت عليها مسبقاً (M & M.N, 2015). باستخدام مقاييس مختلفة لتقييم الأداء (Machine learning metrics)، يجب أن نكون قادرين على تحسين القدرة التنبؤية الشاملة لنموذجنا قبل طرحه للإنتاج على بيانات غير مرئية. أي أنه بدون إجراء تقييم مناسب لنموذج التعلم الآلي باستخدام مقاييس تقييم مختلفة، والإعتماد فقط على الدقة، يمكن أن يؤدي إلى مشكلة عند نشر النموذج المعني على بيانات غير مرئية وقد ينتهي بتوقعات سيئة.

يوجد الكثير من مقاييس التقييم وتختلف حسب المشكلة إذا كانت تنتمي للتعلم بإشراف أو من دون إشراف، وأيضاً قد نختار معيار التقييم حسب نوع المشكلة أي إذا كانت تصنيف أم تنبؤ وما إلى ذلك.

قبل أن نقوم بذكر أهم معايير التقييم، لابدّ لنا أن نذكر مصطلح هام جداً وهذا المصطلح سنكتشف من خلاله معايير التقييم، وهو مصفوفة الارتباك (confusion matrix)، حيث هي عبارة عن جدول يُستخدم في مشاكل التصنيف لتقييم مكان حدوث الأخطاء في النموذج (model)، صفوف هذا الجدول هي الفئات الفعلية التي

ينبغي أن تكون هي النتائج (actual values)، بينما تمثل الأعمدة التوقعات التي قمنا بها (predicted values) (M & M.N, 2015).

وتحتوي على أربعة أمور هامة تساعدنا على فهم الصنف بشكل واضح وهي:

- إيجابي حقيقي TP (True Positive): يشير إلى عدد التنبؤات حيث يتنبأ المصنف بشكل صحيح بالفئة الإيجابية على أنها إيجابية.
- سلبي حقيقي TN (True Negative): يشير إلى عدد التنبؤات حيث يتنبأ المصنف بشكل صحيح بالفئة السلبية على أنها سلبية.
- الإيجابي الكاذب FP (False Positive): يشير إلى عدد التنبؤات حيث يتنبأ المصنف بشكل غير صحيح بالفئة السلبية على أنها إيجابية.
- سلبي كاذب FN (False Negative): يشير إلى عدد التنبؤات حيث يتنبأ المصنف بشكل غير صحيح بالفئة الإيجابية على أنها سلبية. ويوضح الشكل 4 مصفوفة الارتباك (confusion matrix).

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

الشكل 4 مصفوفة الارتباك (confusion matrix) (Vujović, 2021)

ومن خلال هذا الجدول نحصل على معادلات ومفاهيم معايير دقة الخوارزميات وأهمها:

1. **الدقة (Accuracy):** تقيس الدقة ببساطة عدد المرات التي يتوقعها المصنف بشكل صحيح. ويمكننا

تعريف الدقة على أنها نسبة عدد التنبؤات الصحيحة والعدد الإجمالي للتنبؤات.

ومن الجدير بالذكر أن الدقة مفيدة عندما تكون الفئة المستهدفة متوازنة جيدًا ولكنها ليست خيارًا جيدًا

للفئات غير المتوازنة. تخيل السيناريو حيث كان لدينا 99 صورة للكلب وصورة واحدة فقط لقط

موجودة في بيانات التدريب الخاصة بنا. عندها سيتنبأ نموذجنا دائمًا بالكلب، وبالتالي حصلنا على

دقة 99%. في الواقع، دائمًا ما تكون البيانات غير متوازنة، على سبيل المثال البريد الإلكتروني

العشوائي والاحتيال على بطاقات الائتمان والتشخيص الطبي. ومن ثم، إذا أردنا إجراء تقييم أفضل

للمنموذج والحصول على صورة كاملة لتقييم النموذج، فيجب أيضًا مراعاة المقاييس الأخرى

(Vujović, 2021). ويتم حساب الدقة (Accuracy) حسب المعادلة التالية:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **الدقة (Precision):** عدد الحالات التي تم التنبؤ بها بشكل صحيح والتي تبين أنها إيجابية بالفعل.

الدقة مفيدة في الحالات التي يكون فيها الإيجابي الكاذب مصدر قلق أكبر من السلبيات الكاذبة.

تكمّن أهمية الدقة في أنظمة التوصية بالموسيقى أو الفيديو، ومواقع التجارة الإلكترونية، وما إلى

ذلك، حيث يمكن أن تؤدي النتائج الخاطئة إلى إزعاج العملاء وقد يكون ذلك ضارًا بالعمل

(Vujović, 2021). وتوضح المعادلة التالية كيفية حسابها:

$$Precision = \frac{TP}{TP + FP}$$

3. **الاستدعاء (Recall):** عدد الحالات الإيجابية الفعلية التي تمكنا من التنبؤ بها بشكل صحيح

باستخدام نموذجنا. إنه مقياس مفيد في الحالات التي يكون فيها السلبي الكاذب مصدر قلق أكبر من

الإيجابي الكاذب (Vujović, 2021). ويمكن حسابها حسب المعادلة التالية:

$$Recall = \frac{TP}{TP + FN}$$

4. **F1_Score:** يعطي فكرة مجمعة حول مقاييس الدقة (Precision) والاستدعاء (Recall). وهو الحد

الأقصى عند تساويهما. ويمكن أن تكون درجة F1 مقياس تقييم فعال في الحالات التالية:

- عندما تكون تكلفة FP و FN متساوية.
- لا تؤدي إضافة المزيد من البيانات إلى تغيير النتيجة بشكل فعال
- السلبي الحقيقي مرتفع

ويمكن حسابه عبر المعادلة التالية:

$$F1_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ولكن لماذا نستخدم عدة معايير للتقييم ولا نكتفي بمعيار واحد؟

نستخدم عدة معايير للتقييم لأن التقييم بمعيار واحد غير كافٍ لرؤية كيفية تصرف المصنّف بمختلف الحالات وكيفية تعامله مع مختلف أنواع البيانات فقد تكون البيانات متوازنة أو غير متوازنة ولذلك إيجاد عدة معايير يساعدنا على فهم المصنّف بشكل أوضح والتأكد من سلامة العمل مع البيانات غير المرئية. والجدير بالذكر أنه يوجد أيضاً معايير أخرى للتقييم غير التي ذكرناها ولكن قمنا بذكر أهم المعايير.

4 الفصل الرابع: تنفيذ المشروع

4.1 البيئة والأدوات المستخدمة

قمنا باستخدام لغة البرمجة python للعمل على هذه المشروع حيث تعتبر لغة بايثون من اللغات الرائدة في تحليل البيانات والتعلم الآلي وخاصة بوجود عدة مكتبات modules قمنا بالتعامل معها وهي:

🌈Pandas: للقيام بالعمليات المناسبة على ال dataset مثل حذف أو اضافة عمود، التعامل مع القيم الفارغة، التعامل مع عمليات التحويل المناسبة للبيانات لتكون بالشكل المطلوب لخوارزميات التعلم الآلي.

🌈Numpy: للتعامل مع المصفوفات، أساس أي خوارزمية تعلم آلي وأي تقنية من هذه التقنيات هو رياضي فلا بد من جعل كل مميزة مصفوفة لتكون دخل إلى خوارزمية التعلم الآلي، ولذلك هذه المكتبة غنية جداً وتساعدنا في العمل في هذا المجال.

🌈Seaborn: مكتبة رائعة لرسم الخرائط الحرارية التي توضح ال visualization بشكل رائع ومفهوم ليثني لنا فهم التحليل الصحيح، وقمنا باستخدامها لرسم ال confusion matrix.

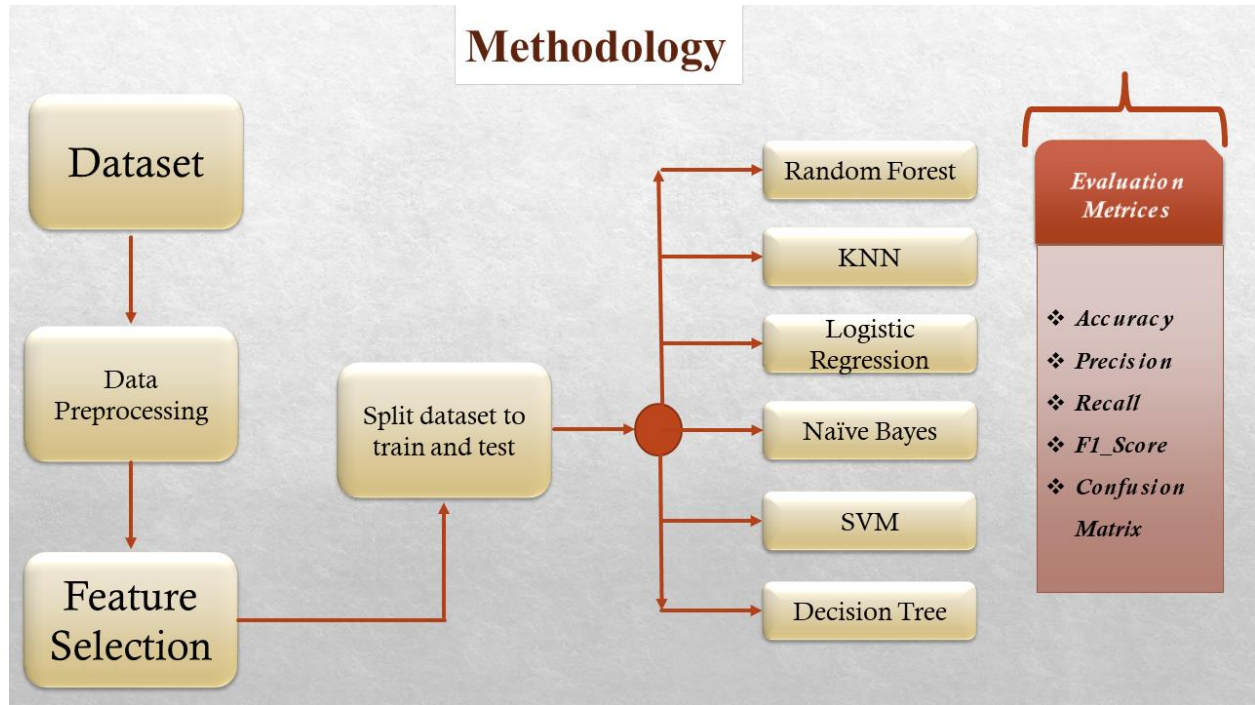
🌈Matplotlib: أشهر مكتبات لغة بايثون لرسم المخططات البيانية، تساعدنا على رسم الدقة والخطأ لكل خوارزمية ويمكن مخطط تفصيلي لمقارنة نتائج الخوارزميات معاً مثلاً.

🌈Scikit learn: مكتبة التعلم الآلي التي تحوي مختلف خوارزميات التعلم الآلي الشهيرة، حيث توفر لنا هذه الخوارزميات لنتمكن من استخدامها وضبطها ببساطة، وأيضاً تحوي جميع معايير التقييم المختلفة للخوارزميات، وخوارزميات اختيار المميزات أيضاً موجودة فيها، حيث جميع الأعمال التي تخص التعلم الآلي قمنا فيها باستخدام هذه ال module.

قمنا باستعمال بيئة google colab التي توفر جميع هذه المكتبات حيث تم تصميمها خصيصاً لأعمال التعلم الآلي والتعلم العميق، هذه البيئة تتيح أحدث اصدار من مختلف هذه ال modules ويتم فيها التدريب على السحابة cloud، كما قمنا بتجربة الكود أيضاً على Jupiter notebook في محرر Anaconda navigator وهي مشابهة تماماً لل google colab حيث يمكنك أن تضع النص text وال code بشكل منظم ورائع يمكنك من شرح خطوات العمل بطريقة مبسطة وواضحة.

4.2 المنهجية

في هذا المشروع قمنا باكتشاف مرض باركنسون باستخدام هذه البيانات المجمعة من تسجيلات صوتية، ومن خلال تطبيق منهجيتنا على هذه البيانات، قمنا بتصنيف الأشخاص إلى أشخاص مصابين وأصحاء، الشكل 5 يوضح المنهجية التي اتبعناها. الدخل input هو المميزات المستخرجة من التسجيلات الصوتية والخروج output هو معرفة الشخص اذا كان مريض بمرض باركنسون أم لا.

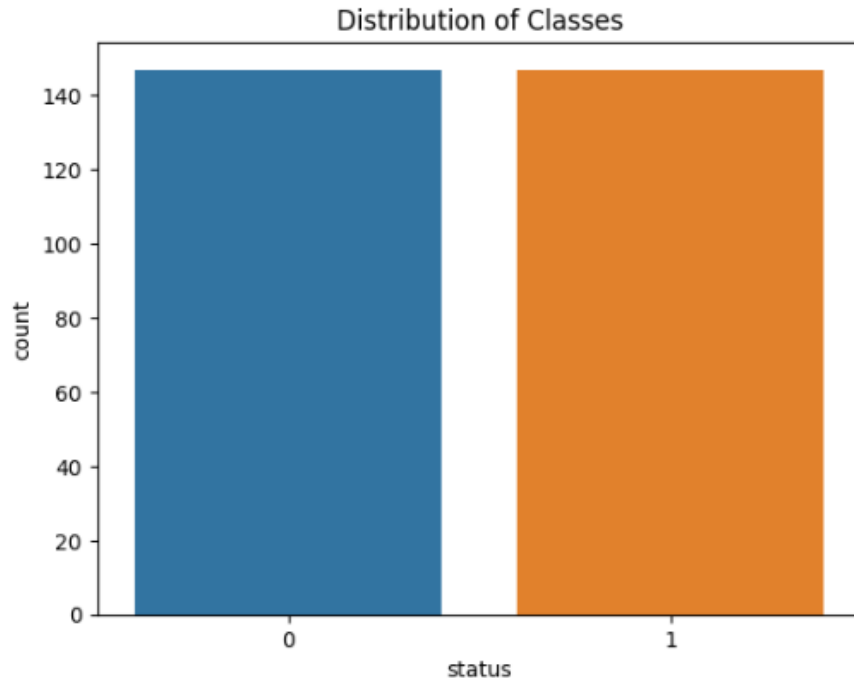


الشكل 5 المنهجية المتبعة في المشروع

في المنهجية التي اتبناها، في البداية نقرأ مجموعة البيانات التي لدينا (تم شرحها مسبقاً في الفصل الثالث)، ثم نقوم بعمليات المعالجة والتحليل على هذه البيانات لتصبح مناسبة وملائمة لتكون دخل لخوارزميات التعلم الآلي، وبعد ذلك نقوم باختيار أفضل المميزات التي سيتم تدريب النماذج عليها وبعد ذلك نقوم بتقييم البيانات لجعل المميزات من نفس المجال، ثم نقوم بتقسيم البيانات إلى تدريب واختبار (80% للتدريب و 20% للاختبار)، بعد ذلك قمنا بتطبيق 6 خوارزميات تعلم آلي مختلفة لمعرفة أي النماذج سيكون الأفضل في تصنيف مرض باركنسون بناءً على مجموعة البيانات المختارة حيث قمنا بتطبيق معايير التقييم المختلفة على هذه الخوارزميات ورسم مصفوفة الارتباك confusion matrix والمقارنة بين هذه النماذج بالنسبة لبيانات الاختبار. سنقوم بشرح كل مرحلة من مراحل المنهجية بالتفصيل.

4.2.1 معالجة البيانات Data Preprocessing

بعد رفع مجموعة البيانات على google colab واستيراد المكتبات اللازمة وقراءة ملف البيانات، تبدأ مرحلة استكشاف وتنظيف البيانات حيث في البداية نقوم باستكشاف نوع البيانات في كل feature للقيام بالعمليات اللازمة على كل منها لتحويلها لرقمية لتكون مناسبة لخوارزميات التعلم الآلي، في البيانات التي لدينا جميعها رقمية باستثناء اسم المريض (name) حيث هذه الميزة ليست مهمة وبالتالي قمنا بحذفها، بعد ذلك قمنا بالتأكد من أن البيانات لا تحوي قيم فارغة ولا تحتاج للمعالجة (no null values)، في الخطوة الأخيرة يجب التأكد من أن البيانات متوازنة، ولكن لدينا عدد العينات للأشخاص الأصحاء أكثر من عدد العينات للأشخاص المصابين بباركنسون وبالتالي يجب علينا موازنة البيانات (balancing dataset) وقمنا بذلك عن طريق زيادة عدد عينات الأشخاص المصابين باستخدام تقنية upsampling وبهذا نحصل على بيانات متوازنة (balanced data). الشكل 6 يوضح البيانات بعد أن أصبحت متوازنة.



الشكل 6 البيانات بعد أن أصبحت متوازنة

4.2.2 اختيار المميزات

في هذه المرحلة قمنا باختيار 10 مميزات من أصل إجمالي عدد المميزات التي سيتم تدريب النموذج عليها، تم ذلك باستخدام chi square test حيث يتم استخدام اختبار مربع كاي (chi square test) في الإحصاء لاختبار استقلالية حدثين. بالنظر إلى بيانات متغيرين، يمكننا الحصول على العدد المرصود O والعدد المتوقع E. يقيس Chi-Square مدى إنحراف العدد المتوقع E والعدد المرصود O لبعضهما البعض. عندما يكون هناك ميزتان مستقلتان، يكون العدد المرصود قريباً من العدد المتوقع، وبالتالي سيكون لدينا قيمة Chi-Square أصغر. تشير قيمة Chi-Square المرتفعة إلى أن فرضية الاستقلال غير صحيحة. بكلمات بسيطة، كلما زادت قيمة Chi-Square، أصبحت الميزة أكثر اعتماداً على الإستجابة ويمكن إختيارها للتدريب النموذجي. هذه المعادلة تعبر عن chi square في اختيار المميزات:

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where: c is degrees of freedom.

O: observed value.

E: expected value.

حيث عدد درجات الحرية يعطى بالعلاقة:

$$\text{degrees of freedom} = (r - 1) * (c - 1)$$

حيث r تعبر عن الأسطر و c تعبر عن الأعمدة.

الشكل 7 يمثل عينة من مجموعة البيانات (المميزات المختارة).

	MDVP:F0(Hz)	MDVP:F1(Hz)	MDVP:Shimmer	MDVP:Shimmer(dB)	MDVP:APQ	Shimmer:DDA	NHR	HNR	spread2	PPE
0	119.992	74.997	0.04374	0.426	0.02971	0.06545	0.02211	21.033	0.266482	0.284654
1	122.400	113.819	0.06134	0.626	0.04368	0.09403	0.01929	19.085	0.335590	0.368674
2	116.682	111.555	0.05233	0.482	0.03590	0.08270	0.01309	20.651	0.311173	0.332634
3	116.676	111.366	0.05492	0.517	0.03772	0.08771	0.01353	20.644	0.334147	0.368975
4	116.014	110.655	0.06425	0.584	0.04465	0.10470	0.01767	19.649	0.234513	0.410335

الشكل 7 عينة من المميزات المختارة

حيث قامت هذه الطريقة باختيار المميزات الموضحة بالشكل 7 بناءً على المعادلة السابقة، وهذا يدل على أنّ هذه المميزات ستؤثر إيجاباً بشكل أكبر على تدريب النموذج من المميزات الأخرى، وبالتالي سنقوم باستخدامها لتدريب نماذج التعلم الآلي لكشف مرض باركنسون.

4.2.3 تقسيم البيانات إلى تدريب واختبار

قمنا بتقسيم البيانات الى تدريب واختبار (80% للتدريب و 20% للاختبار)، قبل التقسيم قمنا بتقييس البيانات (scaling) باستخدام min max scaler حيث يساعد التقييس بجعل المميزات ضمن نفس المجال (بين 0 و1)، وأيضاً يساعد في حل مشكلة ال outliers الموجودة في ال dataset، معادلته كالتالي:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4.2.4 تطبيق خوارزميات التعلم الآلي

بعد تطبيق المراحل السابقة كلّها، أصبحت لدينا البيانات جاهزة للتدريب باستخدام خوارزميات التعلم الآلي المختلفة، قمنا بتطبيق عدة خوارزميات للمقارنة بين النتائج باستخدام معايير التقييم المختلفة واختيار النموذج الأفضل الذي يمكن الإستفادة منه حيث يستطيع الكشف عن مرض باركنسون بدقة وحرفيّة أعلى من النماذج الأخرى، خوارزميات التعلم الآلي المستخدمة هي:

✚ Random Forest Classifier

✚ K-Nearest Neighbors (k = 3)

✚ Logistic Regression

✚ Naïve Bayes

✚ Support Vector Machine

✚ Decision Tree Classifier

قمنا بتدريب هذه الخوارزميات وحساب التقييمات (Accuracy, Precision, Recall, F1_Score)

لكلّ منها وحساب ال confusion matrix ورسم الدقة والخطأ لكل خوارزمية.

مقارنة خوارزميات متعددة لكشف مرض باركنسون يعطي نظرة عامة لقوة وأداء كل خوارزمية في التعامل

مع هذه المشكلة وهذا يعطي ارثاء مفيد للأبحاث القادمة في هذا المجال.

5 الفصل الخامس: الخلاصة والنتائج

في هذا الفصل سنعرض النتائج التي توصلنا إليها في مختلف الخوارزميات ونقارن بينها، وسنقوم بشرح الخلاصة من هذا المشروع والأعمال المستقبلية الممكنة لتطوير هذا المشروع للوصول إلى نموذج قادر على كشف مرض باركنسون بدقة عالية.

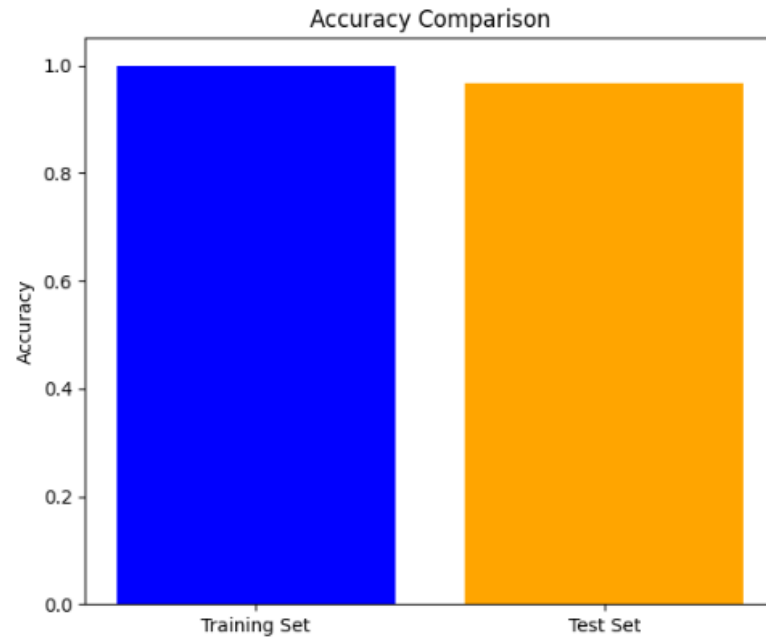
5.1 النتائج

بعد القيام بتطبيق المنهجية التي طرحناها لابدّ من شرح النتائج التي تم الوصول إليها، فيما يلي الجدول يعرض نتائج تقييم الخوارزميات على بيانات الاختبار.

الجدول 2 نتائج تطبيق الخوارزميات على بيانات الاختبار

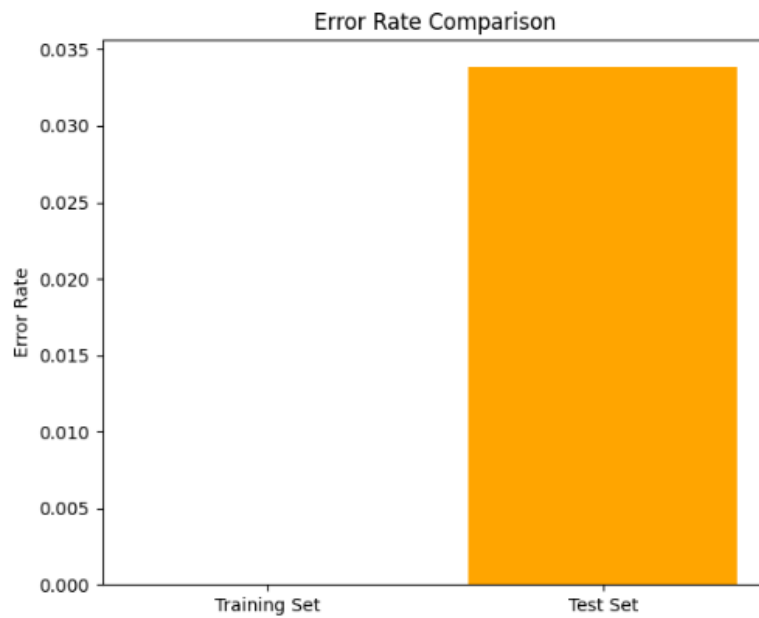
model	Accuracy	Precision	Recall	F1_Score
NB	76.2	95	59.3	73
LR	79.6	91.6	68.7	78.5
SVM	86.4	100	75	85.7
KNN	91.5	100	84.3	91.5
DT	94.9	100	90.6	95
RF	96.6	100	93.7	96.7

من الجدول السابق نلاحظ أن خوارزمية Random Forest هي الأفضل من ناحية مختلف معايير التقييم. الشكل 8 يوضح دقة Accuracy خوارزمية Random Forest.



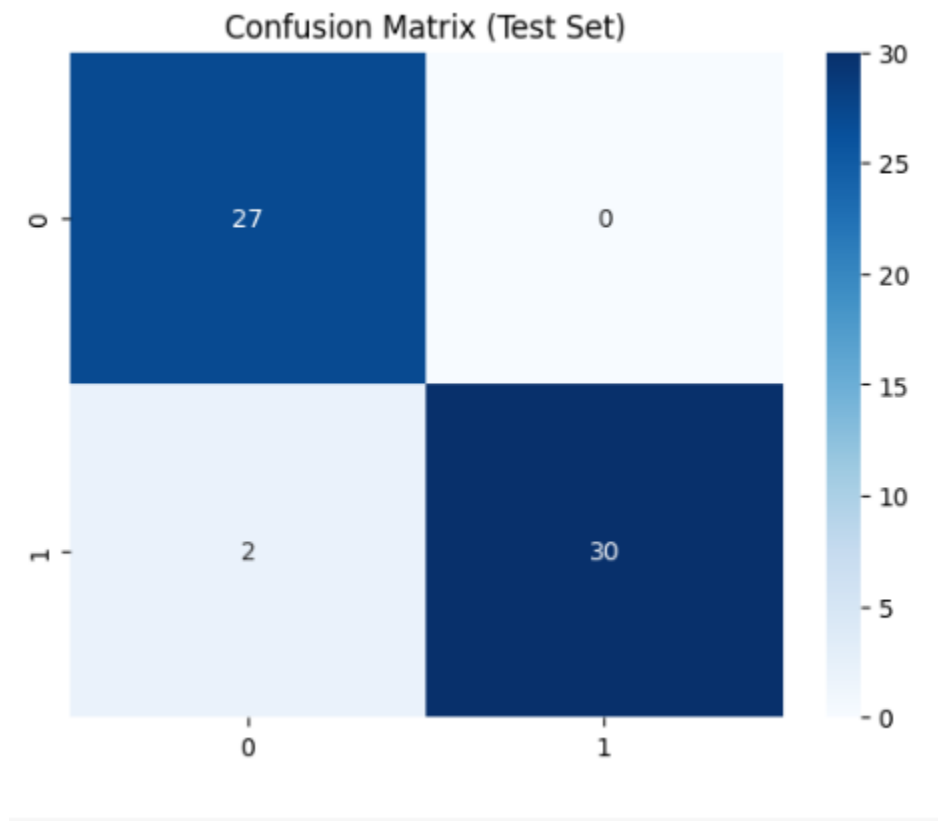
الشكل 8 دقة خوارزمية RF

وأيضاً يوضح الشكل 9 الخطأ لخوارزمية Random Forest على بيانات التدريب والاختبار.



الشكل 9 خطأ خوارزمية RF على بيانات التدريب والاختبار

ولدينا الشكل 10 يوضح مصفوفة confusion matrix لخوارزمية Random Forest.



الشكل 10 مصفوفة confusion matrix لنتائج RF

سنقوم بشرح نتائج هذه المصفوفة (بيانات الاختبار):

القيمة (27) هي عدد التنبؤات الصحيحة للأشخاص الذين ليسوا مصابين بمرض باركنسون (True Negatives).

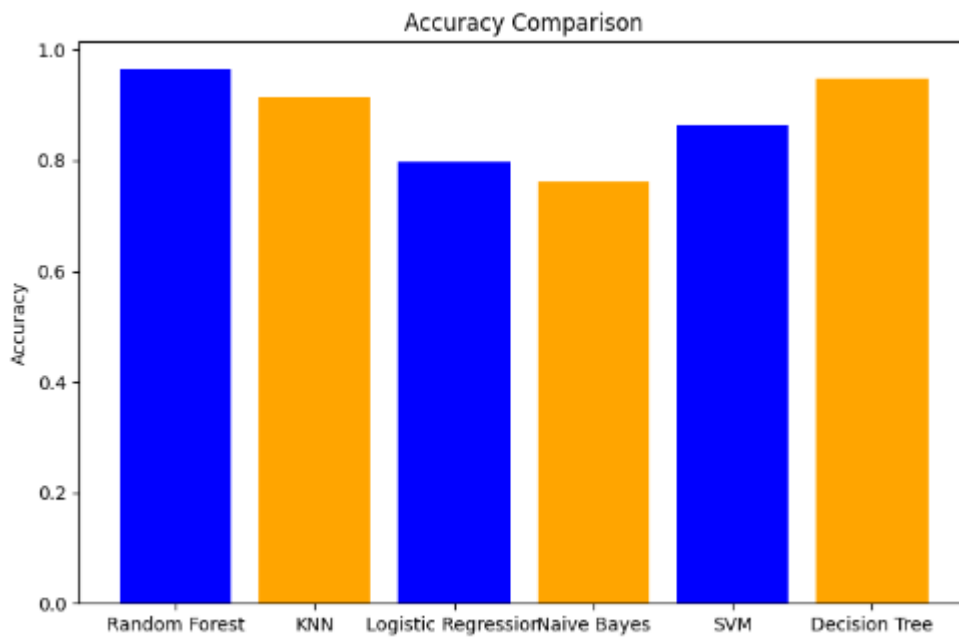
القيمة (30) هي عدد التنبؤات الصحيحة للأشخاص المصابين بمرض باركنسون (True Positives).

القيمة (2) هو عدد التنبؤات الخاطئة حيث تم تصنيف شخص غير مصاب بالمرض كمريض (False Positives).

القيمة (0) تشير إلى أنه لم يتم تصنيف أي شخص مريض على أنه غير مريض (False Negatives).

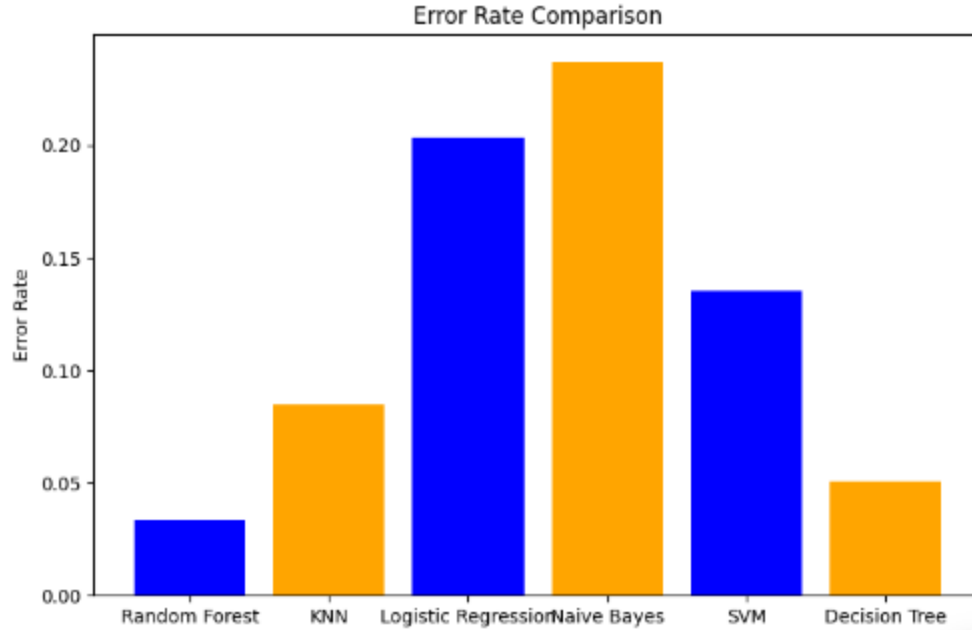
هذه الأرقام تعكس أداء النموذج في التنبؤ بمرض باركنسون. فمن الواضح أن النموذج يتمتع بأداء جيد نسبياً، حيث أنه تمكن من التنبؤ بشكل صحيح بالحالات الإيجابية والسلبية في معظم الأحيان.

لدينا الشكل 11 يوضح المقارنة بالدقة بين مختلف الخوارزميات التي تعاملنا معها.



الشكل 11 المقارنة بالدقة بين مختلف الخوارزميات

والشكل 12 يوضح المقارنة بالخطأ بين مختلف الخوارزميات.



الشكل 12 المقارنة بالخطأ بين مختلف الخوارزميات

5.2 الخلاصة

قمنا في هذا المشروع بالكشف عن مرض باركنسون باستخدام بيانات مستخرجة من تسجيلات صوتية لأشخاص مصابون بالمرض وأشخاص أصحاء، حيث استطعنا التعامل مع هذه البيانات وموازنتها واستخدام عدة نماذج تعلم آلي معها وتقييمها بالنسبة لبيانات الإختبار باستخدام معايير التقييم المختلفة لمعرفة النموذج الأفضل والأسوأ لحل هذه المهمة، حيث أن النموذج الأفضل يمكن وضعه حيز التجربة للقيام بعمليات التحسين عليه. أفضل نتيجة حصلنا عليها كانت مع خوارزمية Random Forest حيث حصلنا على دقة 96.6% ومعدل F1_Score هو 96.7% وهذا يعتبر جيد جداً حيث لم يخطئ النموذج بأي عينة من عينات المرضى ولكن أخطأ فقط بعينتين من الغير مرضى وصنّفهم على أنهم مرضى. لاشك أن الكشف عن مرض باركنسون مهمة حساسة وفي غاية الأهمية، ومنه نستكشف ضرورة القيام بهذه المشاريع باستخدام التقنيات المختلفة وخصوصاً تقنيات الذكاء الاصطناعي التي أصبحت ضرورية في أي تطبيق.

5.3 الأعمال المستقبلية

فيما يتعلق بالأعمال المستقبلية، هناك عدة اتجاهات يمكن استكشافها وتحسينها لتطوير المشروع وزيادة دقة تصنيف مرض باركنسون. بعض الإقتراحات للأعمال المستقبلية تشمل:

✚ **تحسين المعالجة والتحليل الصوتي:** استكشاف تقنيات تحليل صوتي متقدمة لاستخراج مزيد من المعلومات الدقيقة من التسجيلات الصوتية، مما يمكن أن يساهم في تحسين جودة المميزات المستخدمة في تصنيف المرض.

✚ **تنويع المميزات:** استكشاف مزيد من المميزات الصوتية أو تكامل مميزات من مصادر أخرى، مثل البيانات الحيوية أو البيانات الحركية، لتحسين تمثيل الأشخاص المصابين بمرض باركنسون.

✚ **جمع المزيد من البيانات:** من الممكن زيادة كمية البيانات مما يساعد على جعل النموذج يتعلم الانماط المختلفة بشكل أكبر ويتدرّب أكثر، حتى يمكن بعد ذلك استخدام التعلم العميق Deep Learning يكون أفضل.

هذه الاتجاهات تمثل نقاط انطلاق مهمة لتحسين وتطوير المشروع في المستقبل، مما يساهم في تحسين دقة تصنيف مرض باركنسون وجعل النموذج أكثر فعالية في التشخيص.

- Breiman, L. (2001). *Random Forests* (Vol. 45).
- Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. *Procedia Computer Science*, 218, 249–261.
- Guia, M., Silva, R. R., & Bernardino, J. (2019). Comparison of Naive Bayes, support vector machine, decision trees and random forest on sentiment analysis. *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 1, 525–531. <https://doi.org/10.5220/0008364105250531>
- Little, M. A., Mcsharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (n.d.). *Suitability of dysphonia measurements for telemonitoring of Parkinson's disease*.
- M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 01–11. <https://doi.org/10.5121/ijdkp.2015.5201>
- Mašić, F., Đug, M., Nuhić, J., & Kevrić, J. (2018). Detection of Parkinson's Disease by Voice Signal. In *Lecture Notes in Networks and Systems* (Vol. 28). https://doi.org/10.1007/978-3-319-71321-2_90
- Mou, L., Ding, W., & Fernandez-Funez, P. (2020). Open questions on the nature of parkinson's disease: From triggers to spreading pathology. In *Journal of Medical Genetics* (Vol. 57, Issue 2). <https://doi.org/10.1136/jmedgenet-2019-106210>
- Porter, G., & Gupta, J. (2023). Parkinson's Disease. In *Primary Care Occupational Therapy: a Quick Reference Guide*. https://doi.org/10.1007/978-3-031-20882-9_31
- Putra, P., M. H. Pardede, A., & Syahputra, S. (2022). Analisis Metode K-Nearest Neighbour (Knn) Dalam Klasifikasi Data Iris Bunga. *Jurnal Teknik Informatika Kaputama (JTik)*, 6(1).
- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1). <https://doi.org/10.11613/BM.2014.003>
- Sveinbjornsdottir, S. (2016). The clinical symptoms of Parkinson's disease. In *Journal of Neurochemistry*. <https://doi.org/10.1111/jnc.13691>
- Vakili, M., Ghamsari, M., & Rezaei, M. (n.d.). *Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification*.
- Valente, A. X. C. N., Adilbayeva, A., Tokay, T., & Rizvanov, A. (2016). The Universal Non-Neuronal Nature of Parkinson's Disease: A Theory. *Central Asian Journal of Global Health*, 5(1). <https://doi.org/10.5195/cajgh.2016.231>
- Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, 12(6), 599–606. <https://doi.org/10.14569/IJACSA.2021.0120670>

