**Differentiating AI Voice Generator from Real Human Voice**

**Speech Recognition**

Kelompok 7:

* Bryan Mulia - 2602054764
* Hany Wijaya - 2602089742
* Roland Febrian - 2602086904

1. Latar Belakang

Dengan perkembangan teknologi Artificial Intelligence utamanya di dalam pengembangan Deep Fake Sound juga semakin canggih. Teknologi ini bisa mengubah suara seseorang menjadi orang lain dengan tingkat akurasi yang sangat tinggi dengan hasil yang realistis. Tentunya hal ini memiliki dampak positif sekaligus negatif tetapi dampak negatif memiliki kecenderungan lebih tinggi. Perlu adanya suatu usaha preventif untuk mencegah resiko penipuan, disinformasi, penyalahgunaan identitas dan kriminalitas lainnya akibat dari AI voice changer.

Usaha yang dapat dilakukan untuk menekan tingkat penyalahgunaan Deep Fake Voice bisa dilakukan dengan cara yaitu pembuatan model deteksi dan validasi dari keaslian suara yang ada dalam mengecek suatu informasi audio yang tersebar di masyarakat. Hal itulah yang mendasari kami untuk mengembangkan model deteksi AI Voice Changer atau Deep Fake Voice. Tantangan terjadi pada saat model dari voice changer mulai untuk memiliki akurasi yang lebih baik lagi sehingga perlu pengembangan pendeteksian lebih jauh lagi. Dengan adanya pengembangan deteksi ini diharapkan adanya keamanan dari pengguna media sosial untuk tidak mudah percaya akan informasi audio yang belum tervalidasi keamanannya. Hal ini secara tidak langsung menekan kabar bohong yang tersebar di masyarakat dalam bentuk suara. Keamanan dan privasi dari pengguna media sosial juga lebih terjaga karena penyalahgunaan identitas pribadi juga dapat dicegah dengan model ini. Regulasi dari penggunaan AI Voice Changer yang terus dikembangkan akan dipermudah dengan pembuatan model deteksi yang akurat.

1. Rumusan Masalah

Rumusan masalah untuk project kami adalah sebagai berikut:

1. Model apa saja yang dapat secara akurat mendeteksi penggunaan AI Voice Changer atau tidak?
2. Bagaimana cara kerja algoritma dari pengembangan model yang ada untuk mendeteksi penggunaan AI Voice Changer?
3. Faktor apa saja yang menjadi pedoman saat mengklasifikasikan suatu audio masuk ke dalam kategori AI atau tidak?
4. Bagaimana cara evaluasi model yang dikembangkan?
5. Bagaimana rencana kerja kedepannya agar model ini dapat diintegrasikan dalam penerapannya di kehidupan sehari-hari?
6. Tujuan Project

Tujuan pengembangan dari project ini adalah meningkatkan keamanan digital utamanya dalam penerapannya pada sistem keamanan berbasis audio. Selain itu, project ini dikembangkan untuk menekan angka tingkat kriminalitas dalam penyalahgunaan data pribadi menggunakan audio yang ada. Model ini juga dikembangkan untuk membantu validasi terhadap berita atau informasi yang berbentuk audio sehingga mencegah terjadinya penyebaran berita bohong di masyarakat. Model yang kami kembangkan juga memiliki maksud menghasilkan akurasi yang tinggi sehingga kevalidan data audio yang ada semakin terpercaya. Terakhir, tujuan dari pengembangan model ini adalah diharapkan membantu pembuatan regulasi dalam penggunaan AI Voice Changer dan penegakannya sesuai regulasi yang ada.

1. Ruang Lingkup Project

Dalam pengembangan model, kami berfokus kepada pengembangan model deteksi suara AI Voice Changer. Pengembangan menggunakan pendekatan 3 jenis model yaitu KNN, Naive Bayes, dan SVC. Dalam melakukan training dan testing kami menggunakan dataset secara terbatas untuk memastikan keberhasilan dari model yang kami buat dengan labeling audio real dan fake. Keterbatasan bahasa dan dialek juga masih terjadi diakibatkan pengembangan model ini akan berfokus hanya pada penggunaan bahasa Inggris dengan jumlah orang yang dilibatkan juga masih terbatas. Model berfokus dengan membandingkan intonasi, pitch, dan timbre suara dari pemilik suara dalam melakukan validasi. Model kami masih berfokus pengembangannya dalam lingkup pengembangan fitur dan belum pada penerapan pada integrasi dengan fitur lainnya.

1. Deskripsi Dataset

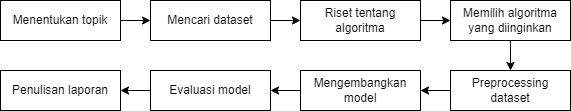
Dataset yang kami gunakan berasal dari *website* Kaggle, yang meliputi suara dari tokoh-tokoh publik yang telah dikonversikan ke beberapa suara berdasarkan tokoh-tokoh publik lainnya.

Berikut link dari datasetnya: <https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition>

Untuk data audionya dibagi menjadi 2 directory, FAKE dan REAL. Untuk yang REAL, datanya mencakup 8 file audio original dari masing-masing tokoh publik. Untuk yang FAKE, datanya mencakup 56 file audio yang telah dikonversikan ke masing-masing suara tokoh yang lain.

1. Tahap-tahap Eksperimen

Diagram tahap eksperimen yang kami lakukan dapat dilihat dari diagram berikut:



Tahap pertama yang kami lakukan untuk eksperimen ini adalah menentukan topik yang ingin diteliti, yaitu klasifikasi penggunaan voice changer AI. Lalu, kami mencari dataset yang sesuai untuk penelitian ini, yang akhirnya didapatkan dari Kaggle dengan audio suara manusia sebenarnya dan yang sudah mengalami voice changer AI. Setelah menemukan dataset yang ingin digunakan, kami melakukan riset tentang algoritma yang sering digunakan untuk masalah serupa (klasifikasi suara), termasuk Resnet50, Logistic Regression, Naive Bayes, KNN, SVM, dsb. Setelah itu, kami memilih algoritma yang ingin dibandingkan. Kami memilih Naive Bayes, KNN, dan SVC karena tiga algoritma ini merupakan algoritma machine learning yang cara kerjanya relatif mudah untuk dikuasai, namun ketiga algoritma tersebut memiliki sifat yang berbeda-beda. Naive Bayes mengasumsikan independensi pada fitur, SVC memungkinkan pemilihan parameter yang optimal, dan KNN yang mempertimbangkan fitur secara keseluruhan.

Untuk mengembangkan ketiga model ini, kami pertama melakukan preprocessing kepada dataset kami untuk mempersiapkan data training dan testing bagi pengembangan. Lalu, dengan ini kami mengembangkan ketiga modelnya dan melakukan evaluasi. Terakhir, performa dari ketiga model ini akan dibandingkan dan dianalisa untuk mendapatkan hasil penelitian.

1. Preprocessing

Pertama-tama kita import library dan fungsi sesuai dengan kebutuhan fungsional yang ingin kami gunakan, seperti numpy, os, librosa, dll. Setelah itu, kita membaca file inputan dari data yang sudah di *folder* guna menampilkan *path file* sesuai dengan *directory* yang telah ditentukan. Lalu, kami melakukan deklarasi variabel untuk *dataset\_path* dari directory inputan data dan *augmented\_dataset\_path* sebagai tempat dataset yang sudah dilakukan preprocessing.

Setelah itu, data audio akan kami segmentasikan menjadi sampel-sampel 10 detik dan akan dibuatkan dua folder untuk menyimpan hasilnya. Kemudian, agar data yang digunakan untuk proses pelatihan dan pengujian nanti tidak biased, kami mengambil jumlah sampel data FAKE yang sama dengan jumlah sampel data REAL. Data FAKE yang terpilih ini akan disimpan ke directory baru yang bernama SELECTED\_FAKE.

1. Rancangan Model

Kami membandingkan perancangan model deteksi penggunaan AI Voice Changer dengan 3 algoritma, yaitu KNN, Naive Bayes, dan SVC. Rincian perancangannya adalah sebagai berikut:

1. K-Nearest Neighbors (KNN)

KNN sendiri merupakan model training non-parametrik dan *supervised learning classifier,* yang menggunakan *proximity* untuk classification atau prediksi terkait kumpulan pada masing-masing titik data.

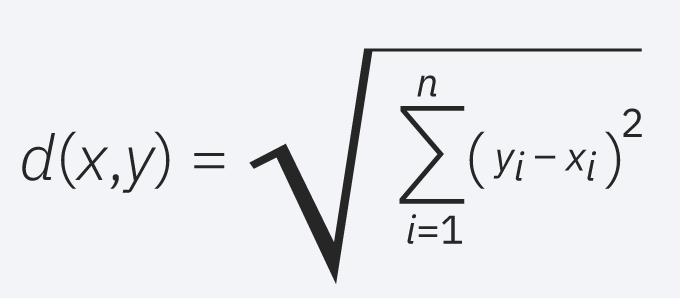


Untuk mengembangkan model KNN, kami perlu memperhatikan *distance metrics* nya yaitu jarak antara titik data satu dengan titik data lainnya dalam menciptakan kumpulan data similaritas dengan batasan-batasan tertentu sebagai *boundaries*.

Berikut beberapa *distance metrics* yang dapat digunakan:

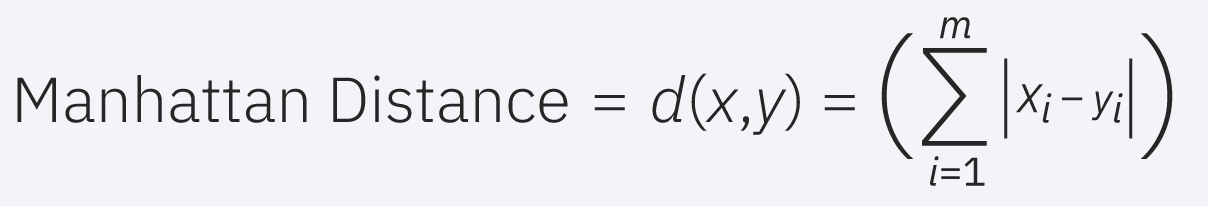
1. Euclidean distance

Teknik ini mengukur jarak antara titik satu dengan titik lainnya secara garis lurus.



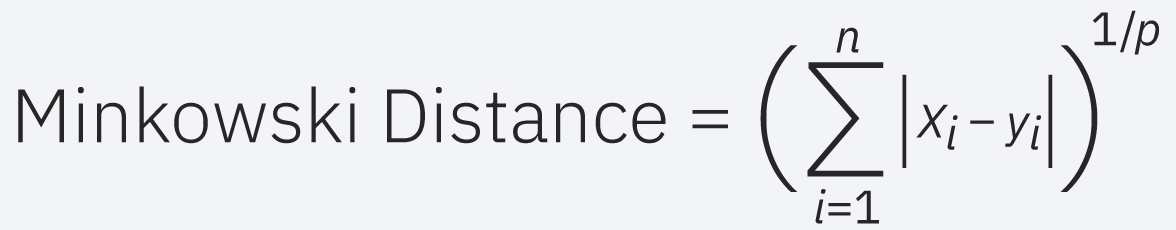
1. Manhattan distance

Teknik ini mengukur jarak antara titik satu dengan titik lainnya dengan meninjau hasil *absolute* atau jumlah dari jarak vektor X dan jarak vektor Y nya untuk mencapai titik lainnya.



1. Minkowski distance

Teknik ini merupakan hasil generalisasi dari Euclidean distance dan Manhattan distance.



dengan nilai p adalah parametrik yang digunakan untuk menentukan jarak dari titik satu dengan titik lainnya.

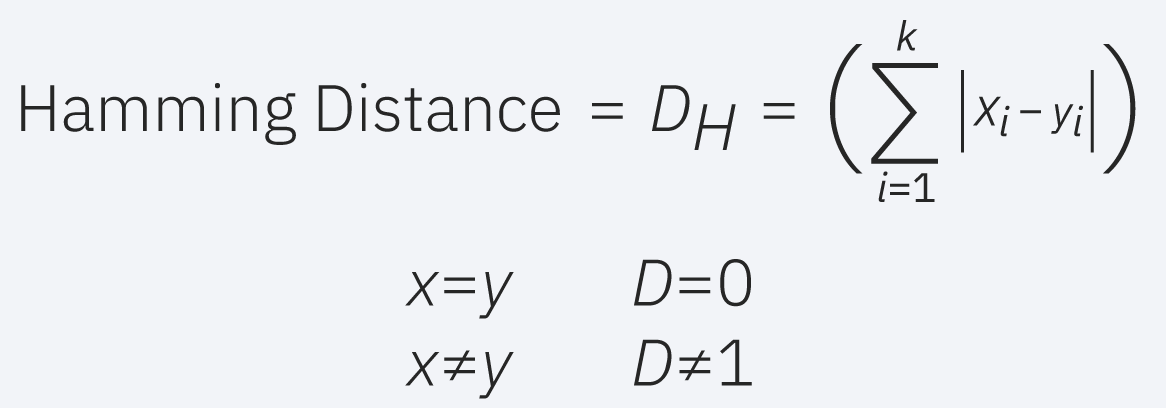
Jika p = 1, maka akan menjadi Euclidean distance.

Jika p = 2, maka akan menjadi Manhattan distance.

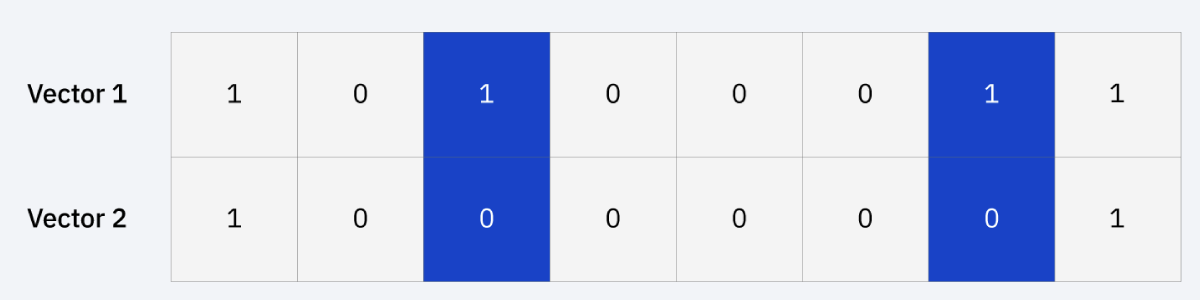
Jika p = , maka akan menjadi Chebyshev Distance.

1. Hamming distance

Teknik ini sering digunakan untuk vektor *Boolean* dan *String* dengan meninjau perbandingan *value* yang berbeda untuk ditambahkan ke nilai *hamming distance*.



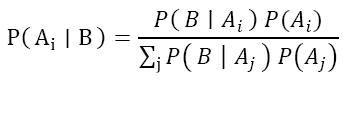
Contoh perhitungan sebagai berikut:



Hamming distance = 1 + 1 = 2

1. Naive Bayes

Naive Bayes sebagai algoritma dalam machine learning merupakan suatu model yang dikembangkan dengan basis teorema Bayes. Naive Bayes mengasumsikan bahwa setiap fitur tidak saling berhubungan atau dapat dikatakan independen satu sama lain. Secara perhitungan, teorema Bayes dinyatakan dengan rumus:

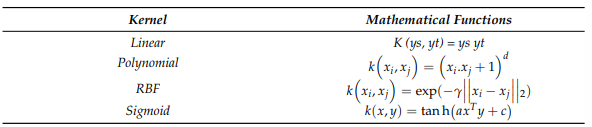


dimana P(A | B) merupakan probabilitas terjadi A jika ada kejadian B, P(B | A) merupakan probabilitas munculnya kejadian B jika terjadi A, dan P(A) merupakan probabilitas terjadi A tanpa memandang kejadian sebelumnya.

Dalam konteks deteksi suara AI, model Naive Bayes akan menghitung probabilitas awal dari masing-masing kelas (suara manusia vs. suara AI) dan probabilitas kondisional dari fitur yang digunakan.

1. Support Vector Classification (SVC)

SVC merupakan cabang pengembangan model dari SVM atau Support Vector Machine. Algoritma dari SVC sendiri berfokus pada pencarian hyperplane (garis dalam suatu bidang) yang memisahkan data yang beragam dengan jarak terdekat antar titik data di tiap kelas yang paling besar. Semakin besar jarak yang ada maka hasilnya akan semakin baik. Penghitungan didasarkan oleh titik yang posisinya paling dekat dengan hyperplane dalam penghitungan yang ada. Terdapat beberapa jenis kernel yang dapat dimanfaatkan yaitu linear, polynomial, Radial Basis Function (RBF), dan sigmoid. SVC efektif untuk melakukan klasifikasi dua kelas yang berbeda yang dalam kasus ini adalah audio real dan fake. Berikut ini adalah penghitungan dari algoritma hyperplane pada SVC:



Dari pengembangan model yang kami lakukan kami memanfaatkan dua jenis kernel yaitu linear dan RBF untuk membandingkan hasil agar akurasinya semakin tinggi. Sebelum masuk dalam tahapan pemanfaatan model, kami memanfaatkan penggunaan K-Fold untuk membagi suatu audio menjadi beberapa bagian dengan pengumpulan fitur menggunakan SelectKBest untuk mendapatkan angka K-Fold yang menghasilkan paling tinggi akurasi yang ada berdasar f-statistik ANOVA dengan pilih fitur 3, 6, 9, 11, 13.

1. Proses Pelatihan Model

Proses pelatihan tiga model kami sebagai berikut:

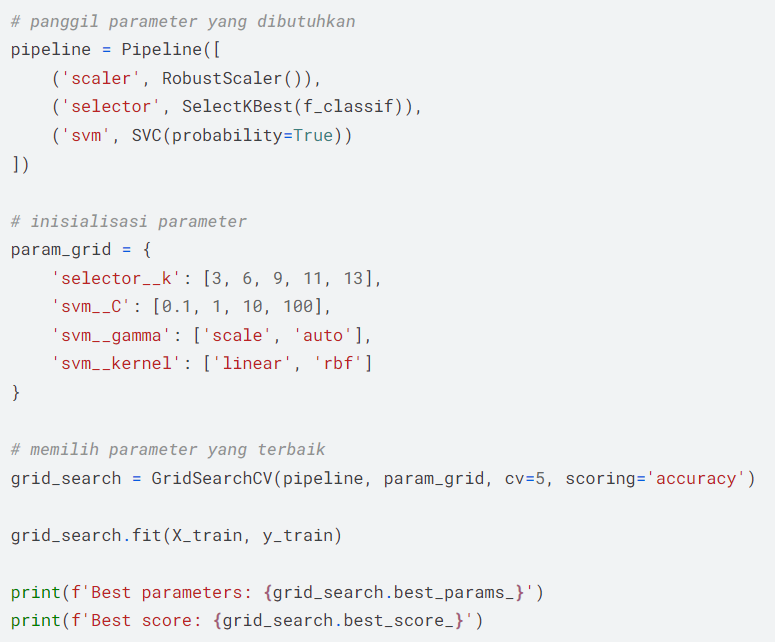
1. K-Nearest Neighbors (KNN)

Dari KNN sendiri meninjau kesamaan dari data testing dan training yang telah akan diproses. Kami menggunakan 2 variabel, yaitu X\_train dan y\_train dari function model selection dari sklearn. Untuk training modelnya, kami melakukan proses *fitting* untuk mengetahui *best fit* nya. Distance metrics yang digunakan merupakan default dari fungsi KNN, yaitu *Euclidean distance.* Kemudian, ada y\_pred untuk mengetahui hasil prediksi dari model KNN terhadap X\_test. Kemudian kami akan mengevaluasi kinerja model dengan menggunakan *accuracy score, precision score, recall score,* dan *recall score*.

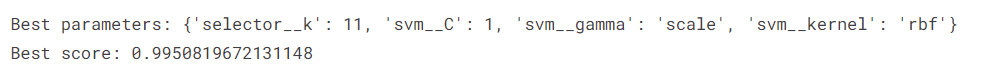
1. Naive Bayes

Untuk melatih model Naive Bayes, kami menggunakan dataset training yang berupa fitur MFCC dari data audio *real* dan *AI voice.* Model akan menghitung probabilitas awal dari kelas data dan probabilitas kondisional dari fitur MFCC yang telah didapatkan. Lalu, dengan teorema Bayes, probabilitas awal dan kondisional akan digunakan untuk menghitung probabilitas akhir yang menyatakan kemungkinan data tersebut ada dalam kelas tertentu. Probabilitas tertinggi akan dijadikan kelas klasifikasi.

1. Support Vector Classification (SVC)



Untuk mengembangkan model SVC, kami memanfaatkan pembentukan model yang ada kami menggunakan function param\_grid() untuk mengeksplorasi hyperparameter dengan berbagai tingkatan value yang berbeda untuk membantu memberikan pilihan parameter untuk menghasilkan akurasi yang tinggi. Function ini terdiri dari jumlah fitur yang dilibatkan, tingkat regularisasi SVC, parameter kernel (scale/auto) dan jenis kernel yang ada (linear/rbf). Setelah itu, kami memanfaatkan fungsi grid\_search() untuk mencari kombinasi hyperparameter dengan hasil akurasi yang terbaik menggunakan cross validation sebanyak 5. Didapatkan informasi seperti berikut untuk parameter dan skor terbaik yang ada:



Dengan hasil parameter ini, kami mengembangkan model SVC yang optimal dan melakukan training terhadap data training yang telah disegmentasi serta menguji model SVC pada data testing untuk memperoleh performnya.

1. Hasil Evaluasi dan Analisis

Dari pengembangan tiga model ini, performa yang didapatkan dengan melakukan pengujian kepada dataset testing sebagai berikut:

|  | **Akurasi** | **Presisi** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| **Naive Bayes** | 0.91 | 0.92 | 0.89 | 0.91 |
| **KNN** | 0.93 | 0.99 | 0.96 | 0.97 |
| **SVC** | 0.99 | 1 | 0.99 | 0.99 |

dan confusion matrix yang didapatkan sebagai berikut:

| Confusion Matrix Naive Bayes | Confusion Matrix KNN |
| --- | --- |
| Confusion Matrix SVC | |

Pengembangan model klasifikasi deteksi suara AI dengan tiga model ini secara keseluruhan dapat dikatakan memiliki performa yang baik, dengan F1-Score yang berada di atas 0.90 bagi ketiga model. Namun, dari ketiga model, Naive Bayes memiliki performa terendah. Hal ini mungkin disebabkan oleh ketidaktepatan penerapan untuk kasus deteksi suara AI karena fitur MFCC yang seringkali memiliki hubungan satu sama lain. SVC merupakan model yang paling unggul di antara ketiganya, dengan akurasi, presisi, recall, dan F1-score yang hampir sempurna, serta hasil confusion matrix yang menunjukkan kesalahan klasifikasi paling sedikit. Model SVC menunjukan kemampuan yang tinggi dalam mengurangi *false positives*. KNN juga menunjukkan performa yang konsisten dengan akurasi, presisi, recall, dan F1-Score yang seluruhnya di atas 0.93.

1. Kesimpulan dan Saran

Dari analisis yang dilakukan, dapat disimpulkan bahwa model Support Vector Classifier (SVC) menunjukkan performa terbaik dalam mendeteksi suara AI, dengan akurasi, presisi, recall, dan F1-score yang hampir sempurna. Model KNN juga memberikan hasil yang sangat baik, meskipun sedikit di bawah SVC, sementara Naive Bayes, meskipun cepat dan sederhana, memiliki performa yang lebih rendah dibandingkan kedua model lainnya. Oleh karena itu, jika prioritas utama adalah akurasi dan minimnya kesalahan klasifikasi, SVC adalah pilihan yang paling tepat, meskipun dengan kebutuhan komputasi yang lebih tinggi.

Kedepannya, terdapat tindakan-tindakan yang dapat dilakukan untuk meningkatkan hasil penelitian, termasuk mengumpulkan lebih banyak data dari kondisi rekaman yang berbeda agar dapat membantu model untuk generalisasi lebih baik serta menggunakan teknik preprocessing suara yang lebih canggih untuk meningkatkan kualitas fitur yang diekstraksi.

1. Lampiran

Berikut kami lampirkan link ke Kaggle Notebook kami: <https://www.kaggle.com/code/emilyheart/deepvoice-ai-detection-ver2/notebook>