Available *online* at : http://jitce.fti.unand.ac.id/
JITCE (Journal of Information Technology and Computer Engineering)
| ISSN (Online) 2599-1663 |

Metode Kernel Distance Classifier Terhadap Klasifikasi Penyakit Jantung

Kasiful Aprianto

Badan Pusat Statistik, Jl Dr Sutomo 6-8, Jakarta, Indonesia

ARTICLE INFORMATION

Received: December 17th , 2021
Revised: May 27th, 2021
Available online: September 30th, 2021

KEYWORDS

Gaussian, Kernel, Support Vector Machine, Kernel Distance Classifier, Classification

CORRESPONDENCE

E-mail: apriantokasiful@gmail.com

A B S T R A C T

This study compares the Support Vector Machine (SVM) and Kernel Distance Classification (KDC) methods to classify heart disease. SVM works by transforming data into higher dimensions using the kernel and classifying data linearly using a hyperplane. Meanwhile, KDC works by finding points that represent each classification from the data that has been transformed into a higher dimension using the kernel, and the new data is predicted based on the closest distance from the point of each classification. The results show that the accuracy produced by SVM is 81.11%. The accuracy produced by the SVM model is better than that produced by the KDC model of 80.47% with a difference of 0.64%, even though both models use kernel transformation.

PENDAHULUAN

World Health Organization (WHO) mengatakan bahwa terdapat sekitar 17.9 juta jiwa telah meninggal karena penyakit jantung dan kerusakan pembuluh darah di tahun 2016 [1]. Hal ini setara dengan 31 persen dari angka kematian seluruh dunia [1]. Melihat angka yang ditunjukkan dari WHO ini menjadikan penyakit jantung menjadi masalah utama dalam dunia kesehatan. Penyebabnya diantaranya yaitu dimulai dari pola hidup dan pola makan, serta kondisi lingkungan yang tidak sehat.

Sedangkan di Indonesia, berdasarkan data dari Riset Kesehatan Dasar (RISKESDAS) tahun 2018, menunjukkan bahwa terdapat 15 dari 1000 penduduk di Indonesia telah menderita penyakit jantung [2]. Hal ini menunjukkan bahwa pentingnya penanggulangan kasus penyakit jantung ini. Hal ini diperkuat oleh hasil SRS (Survey Sample Registration) pada tahun 2014 bahwa terdapat sekitar 12.9 persen kematian di Indonesia disebabkan oleh penyakit jantung [3].

Beberapa metode telah dilakukan untuk mengetahui apakah seseorang memiliki resiko penyakit jantung. Penelitian sebelumnya telah melakukan percobaan dengan menggunakan metode SVM dan memberikan hasil yang baik ketika dibandingkan dengan beberapa metode lainnya. Penelitian tersebut diantaranya yaitu seperti yang telah dilakukan oleh Tabesh dkk, dimana dengan menggunakan algoritma SVM, akurasi yang dihasilkan untuk klasifikasi penyakit jantung yaitu sebesar 84 persen [4]. Sedangkan penelitian lain yang dilakukan oleh Sandhya, menemukan bahwa SVM dalam melakukan prediksi penyakit jantung memberikan akurasi sebesar 85.97 persen, jaringan syaraf tiruan sebesar 85.30 persen, KNN sebesar 84.12 persen, Naive bayes 81.14 persen, dan RIPPER sebesar 81.08 persen [5]. Melihat hal tersebut, terlihat bahwa SVM memberikan hasil yang lebih baik jika dibandingkan dengan metode lainnya untuk klasifikasi penyakit jantung.

SVM sebenarnya merupakan metode yang bekerja dengan memisahkan kedua klasifikasi dengan sebuah garis yang disebut hyperplane dengan data yang telah ditransformasi ke dimensi yang lebih tinggi seperti yang dijelaskan oleh A. Rodan, dkk [6]. Transformasi data yang dilakukan yaitu dengan menggunakan kernel trick. Dalam penelitian ini dilakukan transformasi data ke dimensi yang lebih tinggi menggunakan kernel trick. Perbedaan yang diusung dalam penelitian ini yaitu pada penggunaan hyperplane. Penelitian ini tidak menggunakan hyperplane, melainkan sebuah titik yang mewakili setiap klasifikasi. Titik tersebut dapat mewakili setiap klasifikasi dengan menggunakan rata-rata atau median dari data dimensi yang lain. Titik inilah yang kemudian digunakan untuk melakukan klasifikasi yang baru dengan melihat titik mana yang paling dekat dengan data real. Secara matematis metode ini jauh lebih sederhana jika dibandingkan dengan SVM. Dengan melihat kesamaan dari cara transformasi data, penelitian ini tidak hanya mencoba mengusung metode alternatif, namun juga membandingkan dengan metode

https://doi.org/10.25077/jitce.5.02.70-74.2021

lain yang memiliki beberapa kesamaan tersebut. Penelitian ini melakukan perbandingan antara SVM dengan metode yang diusung.

Support Vector Machine

Support vector machine (SVM) bekerja dengan memaksimalkan batas garis lurus yang akan memisahkan data terhadap masing-masing klasifikasi. Setiap data latih yang ditunjukkan oleh (x_i,y_i) dengan i=1,2,…,N dan x_i=[x_i1,x_i2,…,x_iq]^T merupakan variable independent dan y_i∈{-1,+1} merupakan variable dependent atau label kelas. Hyperplane klasifikasi untuk model ini dapat ditulis seperti pada persamaan:

$$\mathbf{w}.\mathbf{x_i} + b = 0 \quad (1)$$

Untuk kelas -1 menggunakan persamaan

$$\mathbf{w}.\mathbf{x_a} + b \leq -1 \quad (2)$$

Dan untuk kelas +1 menggunakan persamaan

$$\mathbf{w}.\mathbf{x_b} + b \geq +1 \quad (3)$$

Dengan melihat persamaan (2) dan (3), maka jarak antar klasifikasi dapat dihitung dengan melihat selisih dari batas maksimal dari kelas -1 yaitu $\mathbf{w}.\mathbf{x_a} + b = -1$ atau $\mathbf{w}.\mathbf{x_a} = -1 - b$ dengan batas minimal dari kelas +1 yaitu $\mathbf{w}.\mathbf{x_b} + b = 1$ atau $\mathbf{w}.\mathbf{x_b} = 1 - b$. Selisih tersebut seperti pada persamaan (4):

$$jarak = \frac{|\mathbf{w}.\mathbf{x_b} - \mathbf{w}.\mathbf{x_a}|}{\|\mathbf{w}\|} = \frac{|(\mathbf{1}-\mathbf{b})-(\mathbf{-1}-\mathbf{b})|}{\|\mathbf{w}\|} = \frac{\mathbf{2}}{\|\mathbf{w}\|} \quad (4)$$

Untuk mendapatkan lebar yang optimal, diperlukan langkah untuk meminimalkan nilai $\|\mathbf{w}\|$ atau $\frac{1}{2}.\|\mathbf{w}\|^{\mathbf{2}}$ dengan syarat mengikuti persamaan (2) dan (3) menggunakan Lagrange Multiplier. Persamaan (2) dan (3) dapat dijelaskan dengan membuat persamaan baru seperti yang dijelaskan pada persamaan (5).

$$y_i(\mathbf{w}.\mathbf{x_i} + b) - 1 \geq 0 \quad (5)$$

Dengan: $y = \begin{cases} +1\ jika\ \mathbf{w}.\mathbf{z} + \mathrm{b} > 0 \\ -1\ jika\ \mathbf{w}.\mathbf{z} + \mathrm{b} < 0 \end{cases}$

Sehingga optimasi Lagrange Multiplier seperti pada persamaan (6).

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\mathbf{w}.\mathbf{x_i} + b) - 1 \quad (6)$$

Dengan α adalah Lagrange multiplier yang berkorespondensi dengan x_i

Persamaan (6) untuk selanjutnya diturunkan terhadap $\mathbf{w}$ dan b sehingga

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (7)$$

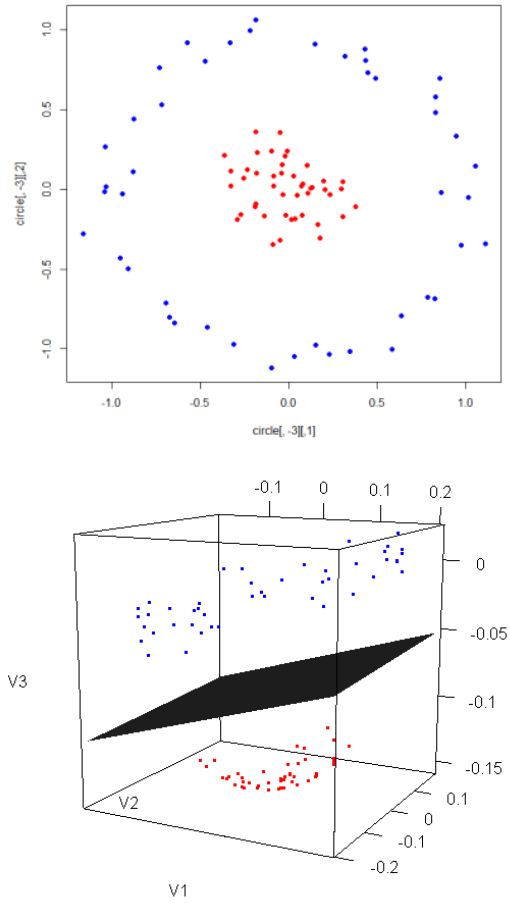
Dan

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (8)$$

Dengan menggunakan persamaan (7) dan (8), maka persamaan (6) dapat ditulis dalam bentuk:

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^{\mathrm{T}} \mathbf{x_j} \quad (9)$$

Nilai α diperoleh dengan menghitung persamaan (9) menggunakan nilai input x dan y, sehingga nilai b juga dapat dihitung menggunakan $b = y_i - w.x_i$ dan hasil prediksi untuk data baru dapat dijelaskan dengan $\sum \alpha_i y_i \mathbf{x_i}^T + b \geq 0$ maka masuk ke kelompok (+). Terlihat juga bahwa pada persamaan (9) terdapat perkalian antara $\mathbf{x_i}^T \mathbf{x_j}$. Disini bisa menggunakan teknik kernel, dimana teknik ini berguna untuk memproyeksikan data dari input space asli ke feature space [7], sehingga memiliki dimensi data yang lebih tinggi dan pada akhirnya data yang tidak bisa dipisah secara linear pada dimensi semula menjadi terpisah pada dimensi yang lain.



Gambar 1. Contoh penerapan SVM

Gaussian Kernel

Banyak teknik data mining atau machine learning yang dikembangkan dengan asumsi linear, sehingga algoritma yang dihasilkan terbatas untuk kasus linear, sedangakan umumnya kasus di dunia nyata adalah kasus yang tidak linear (Santosa). Untuk itu diperlukan suatu cara untuk melakukan transformasi pada data tersebut. Salah satu diantaranya yaitu dengan menggunakan metode kernel (Scholkopf dan Smola, 2002). Dengan metode kernel, data x akan dipetakan kedalam feature space dengan dimensi yang lebih tinggi. Salah satu contoh fungsi kernel adalah kernel Gaussian.

$$k(\bar{x}_i, \bar{x}_j) = \exp\left(-\frac{(\bar{x}_i - \bar{x}_j)^2}{2\sigma^2}\right) \quad (10)$$

dengan $\bar{x}_i$ dan $\bar{x}_j$ adalah vector yang menjelaskan suatu titik dengan dimensi n. Terlihat bahwa $(\bar{x}_i - \bar{x}_j)$ merefleksikan jarak antar titik, dan σ merupakan suatu variable bebas. Misalkan $\boldsymbol{X} = \{\bar{x}_1, \bar{x}_2 \dots \bar{x}_m\}^{\boldsymbol{T}}$ dimana dataset $\boldsymbol{X}$ terdiri dari titik ke-1 sampai titik ke-m, dengan dimensi n vector, maka fungsi $\boldsymbol{K}$ dipetakan seperti berikut:

$$k(\bar{x}_i, \bar{x}_j): \mathbb{R}^n . \mathbb{R}^n \rightarrow \mathbb{R}$$

Fungsi diatas merupakan kernel jika dan hanya jika matriks yang dihasilkan adalah matriks Gram dari inner product. Sebagai bukti, misalkan fungsi kernel untuk titik i dan j dapat dijelaskan dalam rumus berikut:

$$\begin{aligned} k(\bar{x}_i, \bar{x}_j) &= \exp\left(-\frac{(\bar{x}_i - \bar{x}_j)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\{(\bar{x}_i^T . \bar{x}_i) + (\bar{x}_j^T . \bar{x}_j) - 2(\bar{x}_i^T . \bar{x}_j)\}}{2\sigma^2}\right) \\ &= \exp\left(\frac{\bar{x}_i^T . \bar{x}_i}{2\sigma^2}\right) . \exp\left(\frac{\bar{x}_j^T . \bar{x}_j}{2\sigma^2}\right) . \exp\left(\frac{\bar{x}_i^T . \bar{x}_j}{\sigma^2}\right) \end{aligned} \quad (11)$$

Persamaan (11) menunjukkan bahwa dari fungsi yang diberikan, titik i dan j berdimensi n terpetakan ke sebuah koordinat baru dari fungsi $k(x_i, x_j)$. Semua kombinasi dari matriks $\boldsymbol{X'}$ dan $\boldsymbol{X''}$ dapat digambarkan sebagai berikut:

$$\boldsymbol{K}(\boldsymbol{X'}, \boldsymbol{X''}) = \begin{bmatrix} k(\bar{x}'_1, \bar{x}''_1) & \cdots & k(\bar{x}'_1, \bar{x}''_n) \\ \vdots & \ddots & \vdots \\ k(\bar{x}'_n, \bar{x}''_1) & \cdots & k(\bar{x}'_n, \bar{x}''_n) \end{bmatrix} \quad (12)$$

Kernel Distance Classifier (Metode yang Diusulkan)

Kernel Distance Classifier (KDC) bekerja dengan mencari titik yang mampu mewakili keseluruhan data untuk setiap klasifikasi suatu dataset yang telah ditransformasi kedalam bentuk feature space menggunakan kernel. Misalkan $\boldsymbol{X_{train}} = \{\bar{x}_1, \bar{x}_2 \dots \bar{x}_m\}^{\boldsymbol{T}}$ dan $\boldsymbol{Y_{train}} = \{y_1, y_2, \dots, y_m\}$ dengan y_i merupakan klasifikasi dari dataset ke-i dengan kelompok klasifikasi $Q \in \{1,2, \dots, q\}$, dimana dataset $\boldsymbol{X_{train}}$ terdiri dari titik ke-1 sampai titik ke-m, dengan dimensi n vector, maka $\boldsymbol{X_{train}}$ ditransformasikan terlebih dahulu menggunakan kernel

$$\boldsymbol{X_{kernel\_train}} = \boldsymbol{K}(\boldsymbol{X_{train}}, \boldsymbol{X_{train}}) \quad (13)$$

sehingga data yang baru akan menjadi berdimensi m. Selanjutnya koordinat perwakilan dari masing-masing klasifikasi dihitung dengan menggunakan titik pusat dari sekumpulan data setiap klasifikasi. Nilai pusat tersebut bisa menggunakan rata-rata atau nilai tengah. Notasi rata-rata dapat dilihat sebagai berikut:

$$\boldsymbol{T}_{(Q=q)} = \frac{\sum_{i=1}^{n_{(Q=q)}} \bar{x}_{(Q=q)i}}{n_{(Q=q)}} \quad (14)$$

Atau menggunakan nilai tengah

$$\boldsymbol{T}_{(Q=q)} = l + \left(\frac{\frac{n}{2} - cf}{f}\right) h \quad (15)$$

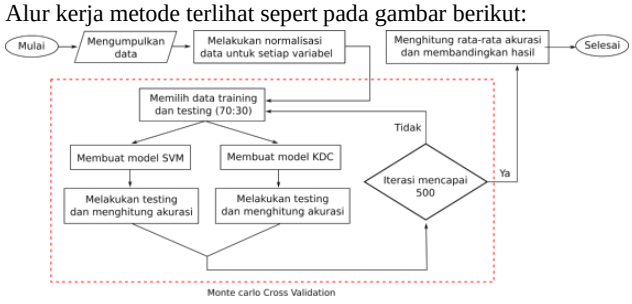
Dengan l adalah nilai terbawah dari kelas median, n adalah jumlah observasi, cf merupakan frekuensi kumulatif, f merupakan frekuensi dari kelas median, serta h merupakan ukuran kelas.

Sedangkan untuk data baru, maka penentuan klasifikasi dapat diketahui dengan melakukan perhitungan jarak ke setiap $\boldsymbol{T}$ dan memilih yang paling dekat dengan titik tersebut. Hal ini dapat dituliskan dengan notasi

$$argmin_Q \left(\sqrt{\left(\boldsymbol{K}(\boldsymbol{X_{test}}, \boldsymbol{X_{train}}) - \boldsymbol{T_q}\right)^2}\right).$$

METHOD

Alur kerja metode terlihat sepert pada gambar berikut:



Gambar 2. Alur Kerja Penelitian

Data yang digunakan untuk penelitian ini yaitu menggunakan data heart disease yang bersumber dari UCI Machine learning [8]. Data terdiri dari 165 baris pasien terkena jantung dan 138 baris pasien tidak terkena jantung. Data ini terdiri dari 14 variabel, dan masing-masing variable dilakukan normalisasi data dengan rumus:

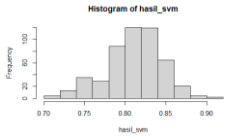
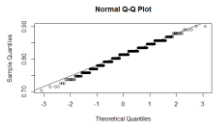
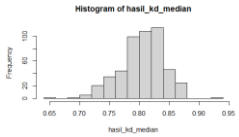
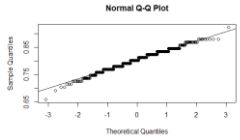
$$x_{normalisasi} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Dimana x_{min} merupakan nilai minimal yang ada pada variable dan x_{max} merupakan nilai maksimal pada variable.

Pengujian metode dilakukan dengan pendekatan monte carlo cross validation, dimana metode ini bekerja dengan melakukan perulangan dengan mengambil sampel training dan testing, lalu menerapkan analisis statistic untuk menghitung hasilnya [9]. Pengujian ini menggunakan 500 kali percobaan dan mendapatkan rata-rata akurasi dari setiap percobaan yang dilakukan oleh setiap metode. Jumlah data training dan testing diambil secara acak dengan proporsi 70:30 untuk setiap perulangan. Pada akhirnya, data yang diperoleh dari setiap perulangan diuji dengan model SVM dan KDC.

https://doi.org/10.25077/jitce.5.02.70-74.2021

TABEL 1. UJI NORMALITAS

Model	Histogram	QQ Plot	Shapiro test p-value
SVM			0.0003617 (p<0.05) Tolak H0
KDC-Median			6.42e-05 (p<0.05) Tolak H0

HASIL DAN PEMBAHASAN

Dari 500 percobaan yang dilakukan, untuk metode SVM akurasinya sebesar 81.11% dengan SD sebesar 3,74%. Sedangkan jika menggunakan metode KDC diperoleh akurasi sebesar 78.05% jika menggunakan rata-rata, dan 80.47% jika menggunakan median, dengan standar deviasi masing-masing 3.62% dan 4.59%. Hasil dapat dilihat pada tabel 1:

TABEL 2. AKURASI UNTUK SETIAP MODEL

Model	Accuracy (%)	Standard Deviation	Confidence Interval (95%)
SVM	81.11	3.47	80.89 < X < 81.32
KDC-mean	78.05	4.59	77.76 < X < 78.33
KDC-median	80.47	3.62	80.24 < X < 80.69

Dari hasil diperoleh, terlihat pada table 1 bahwa model SVM masih memberikan hasil pemodelan yang lebih baik dibandingkan dengan metode KDC baik menggunakan mean maupun median. Selisih antara akurasi yang dihasilkan oleh SVM dengan KDC-median yaitu sebesar 0.64%. Meskipun SVM memberikan hasil yang paling baik dalam pemodelan klasifikasi menggunakan data klasifikasi jantung [8], namun selisih perbedaan akurasi dengan menggunakan KDC-median tidak terlalu jauh. Namun hal ini perlu diuji dengan menggunakan kaidah statistik. Uji yang digunakan yaitu dengan menggunakan Wilcoxon test. Wilcoxon test dipilih karena uji yang dilakukan yaitu membandingkan kedua data yang tidak berdistribusi normal [10].

Dari Tabel 2, disimpulkan bahwa sebaran kedua data berdistribusi tidak normal, terlihat dari uji normalitas menggunakan shapiro wilk dimana nilai p lebih kecil dari 0.05. Dengan demikian, hasil dari uji Wilcoxon test dapat dilakukan. Nilai p dari hasil uji Wilcoxon yaitu 4.76e-05, yang artinya terdapat perbedaan signifikan antara akurasi model SVM dengan KDC-Median, meskipun selisih rata-rata dari kedua model hanya sebesar 0.64%.

KESIMPULAN

Percobaan metode yang dilakukan menunjukkan bahwa metode KDC dengan pemusatan menggunakan median memberikan hasil yang berbeda secara signifikan dengan metode SVM untuk melakukan klasifikasi penyakit jantung. Untuk kasus ini, SVM lebih cocok digunakan karena mampu memberikan nilai akurasi yang lebih baik dibandingkan dengan metode KDC. Namum hasil yang ditemui belum bisa menggeneralisir bahwa metode KDC tidak lebih baik daripada SVM karena belum dilakukan percobaan dengan data yang lain, sehingga masih dibutuhkan kajian yang lebih mendalam dengan menggunakan dataset yang lebih beragam. Selain itu, metode KDC perlu dilakukan kajian lagi dengan mengganti cara menghitung jarak terdekat dari masing-masing perwakilan titik setiap klasifikasi. Salah satunya seperti mempertimbangkan peluang seberapa dekat suatu titik dengan masing-masing titik klasifikasi yang lain.

DAFTAR PUSTAKA

[1] "Cardiovascular diseases (CVDs)." https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Nov. 29, 2020).

[2] "Press Release, World Heart Day PERKI 2019 - News & Event | Perhimpunan Dokter Spesialis Kardiovaskuler Indonesia (PERKI)." http://www.inaheart.org/news_and_events/news/2019/9/26/press_release_world_heart_day_perki_2019 (accessed Nov. 29, 2020).

[3] "Kementerian Kesehatan Republik Indonesia." https://www.kemkes.go.id/article/view/17073100005/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-.html (accessed Nov. 29, 2020).

[4] P. Tabesh, G. Lim, S. Khator, and C. Dacso, "A support vector machine approach for predicting heart conditions," IIE Annu. Conf. Expo 2010 Proc., 2010.

[5] Y. a Sandhy, "Prediction of Heart Diseases using Support Vector Machine," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 2, pp. 126–135, 2020, doi: 10.22214/ijraset.2020.2021.

[6] A. Rodan, H. Faris, J. Alsakran, and O. Al-Kadi, "A Support Vector Machine approach for churn prediction in telecom industry," Inf., vol. 17, no. 8, pp. 3961–3970, 2014.

[7] D. Chen, Q. He, and X. Wang, "On linear separability of data sets in feature space," Neurocomputing, vol. 70, no. 13–15, pp. 2441–2448, 2007, doi: 10.1016/j.neucom.2006.12.002.

[8] "UCI Machine Learning Repository: Heart Disease Data Set." https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Nov. 29, 2020).

[9] Q. S. Xu and Y. Z. Liang, "Monte Carlo cross validation," Chemom. Intell. Lab. Syst., vol. 56, no. 1, pp. 1–11, 2001, doi: 10.1016/S0169-7439(00)00122-2.

[10] Z. Kurucova, J. Medová, and A. Tirpakova, "Cogent Education The effect of different online education modes on the English language learning of media studies students The effect of different online education modes on the English language learning of media studies students," Cogent Arts Humanit., 2017, doi: 10.1080/2331186X.2018.1523514.