مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(١٢) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (١٢) 2023

منهجية لتحسين أداء التعلم الموحد القائم على العنقدة في حوسبة الحافة ذات البيانات غير المتجانسة

د. م. ماهر ابراهيم*

م. بتول علي**

(تاريخ الإيداع ٣/8/ 2023 – تاريخ النشر7/12/ 2023)

□ ملخّص □

يعرف التعلم الموحد FL (Federated Learning) بأنه أسلوب تدريب نماذج التعلم الآلي باستخدام البيانات في البيئات الموزعة التي تعتمد في عملها على المعالجة الموزعة بدلاً من الحاجة إلى خادم مركزي لتخزين ومعالجة البيانات، وهو ما يعرف ب حوسبة الحافة أو الحوسبة الطرفية (Edge Computing)، حيث تجري عملية المعالجة والتخزين على حافة الشبكة. يسمح التعلم الموحد لمجموعة من الأجهزة بتدريب نموذج تعلم آلي محدد على بيانات هذه الأجهزة بشكل متعاون مع الحفاظ على خصوصية البيانات.

تظهر الحاجة إلى إيجاد طرق للتغلب على مشكلة البيانات غير المتجانسة لدى الأجهزة المشاركة في التعلم الموحد لما لها من تأثير على تقارب النماذج المستخدمة وجودة التدريب ودقة النتائج.

تعاني طرق التعلم الموحد التقليدية المعتمدة على تجميع الأجهزة ذات البيانات المتشابهة كمرحلة أولية لعملية التدريب من الحاجة إلى تحديد عدد مجموعات(عناقيد) ثابتة، بالإضافة إلى عدم فعالية نتائج عملية تجميع الأجهزة ذات البيانات المتشابهة.

تم في هذا البحث اقتراح طريقة لتحسين أداء النماذج المدربة باستخدام التعلم الموحد القائم على التجميع من حيث استخدام طريقة جديدة لتحديد تشابه بيانات الأجهزة دون الحاجة لتحديد عدد عناقيد ثابت، مما يزيد من ديناميكية عملية التدريب بالإضافة إلى تقليل عدد البارامترات التي يجب تحديدها مسبقاً وهو ما تفتقر له الدراسات السابقة، وبالتالي زيادة ديناميكية عملية التدريب في التعلم الموحد وتكيفه مع بيئة الحوسبة الطرفية من حيث توافر الأجهزة وتغير المشاركين في التدريب. كما تم اختبار المنهجية المقترحة باستخدام خوارزميتي عنقدة ديناميكية على مجموعة البيانات المعيارية CIFAR 10 مع حالات بيانات مختلفة وأظهرت النتائج تحسناً ملحوظاً في دقة النماذج وتخفيضاً كبيراً بعدد دورات الاتصال اللازمة بين الخادم المركزي والأجهزة الموزعة من أجل الوصول إلى أداء مناسب لنموذج التعلم الآلي.

الكلمات المفتاحية: التعلم الموحد، عدم تجانس البيانات، العنقدة، حوسبة الحافة.

*أستاذ مساعد في قسم هندسة تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس – سوريا.

** مهندسة في قسم هندسة تكنولوجيا المعلومات–كلية هندسة تكنولوجيا المعلومات والاتصالات–جامعة طرطوس–سوريا.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(١٢) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (12) 2023

A methodology for improving the performance of cluster-based federated learning in edge computing with heterogeneous data

Dr. Maher Ibrahim *
Batool Ali **

(**Received 3/8/2023.Accepted 7/12/2023)**

□**ABSTRACT** □

Federated Learning (FL) is defined as a method of training machine learning models using data in distributed environments that rely on distributed processing instead of the need for a central server to store and process data, which is known as edge computing. The processing and storage process take place at the edge of the network. Federated learning allows a group of devices to collaboratively train a specific machine learning model on the data of these devices while maintaining data privacy.

There is a need to find ways to overcome the problem of heterogeneous data on devices involved in federated learning because it has an impact on the convergence of the models used, the quality of training, and the accuracy of the results.

Traditional federated learning methods that rely on grouping devices with similar data as the initial stage of the training process suffer from the need to determine a fixed number of clusters, in addition to the ineffectiveness of the results of the process of grouping devices with similar data.

In this research, a method was proposed to improve the performance of models trained using clustering-based federated learning in terms of using a new method to determine the similarity of device data without the need to specify a fixed number of clusters, which increases the dynamism of the training process in addition to reducing the number of parameters that must be specified in advance, which previous studies is lacking from. Thus, increasing the dynamism of the training process in federated learning and its adaptation to the edge computing environment in terms of availability of devices and the change of training participants. The proposed methodology was also tested using two dynamic clustering algorithms on the CIFAR 10 benchmark dataset with different data cases, and the results showed a significant improvement in the accuracy of the models and a significant reduction in the number of communication cycles required between the central server and distributed devices in order to reach adequate performance for the machine learning model.

Key words: Federated learning, data heterogeneity, clustering, edge computing.

* Associate Professor, Department of Information Technology Engineering, Information and Communication Technology Engineering, Tartous University, Syria.
** Engineer, Department of Information Technology Engineering, Information and Communication Technology Engineering, Tartous University, Syria.

١- مقدمة:

أدت الزيادة في الأجهزة الذكية وإنترنت الأشياء إلى تضخم في حجم البيانات التي يتم توليدها يومياً. وبالتالي ظهرت الحوسبة الطرفية كحل أكثر تناسباً مع طبيعة الشبكات بالمقارنة مع إرسال هذه البيانات الضخمة عبر الشبكة إلى المخدمات السحابية وذلك بسبب قيود عرض الحزمة والخصوصية.

ترافقت الحوسبة الطرفية مع وجود قيود على مشاركة البيانات ومشاكل الخصوصية، بالتالي تم تقديم منهجيات تدريب ذات قدرة على معالجة البيانات المخزنة في الأجهزة الطرفية مع الالتزام بقوانين وسياسات حماية الخصوصية[1] .

يعتبر التعلم الموحد أحد آليات التدريب الموزع الجديدة لنماذج التعلم الآلي حيث يعتمد على التدريب التعاوني لنموذج مشترك باستخدام بيانات المستخدمين الموزعين المشاركين في التدريب دون تبادل هذه البيانات.

على الرغم من محاسن التعلم الموحد إلا أن تطبيقاته في البيئات العملية ما تزال تواجه بعض التحديات ومنها وجود بيانات غير متجانسة عند المستخدمين المشاركين في التدريب أو ما يعرف بعدم التجانس الإحصائي (بيانات المستخدمين ذات توزيعات احتمالية مختلفة). وهو ما ينتج عن عدة عوامل كطبيعة التوزع الجغرافي للأجهزة (مناطق زمنية مختلفة– أحداث مختلفة...) وسلوك المستخدمين. هذه الاختلافات في التوزيعات الاحتمالية تناقض الافتراض الذي تقوم عليه عملية تدرب نماذج التعلم الآلي، حيث تبنى عملية التدريب على افتراض أن كل عينات التدريب مستقلة وتأتي من نفس توزع البيانات (identically distributed).[2]

وبناء على الدراسات السابقة في هذا المجال يعد التعلم الموحد القائم على عنقدة المستخدمين (تجميعهم حسب تشابه توزيعات بياناتهم من أجل التدريب) طريقة تدريب جديدة لنماذج التعلم الآلي للتعامل مع تحديات البيانات غير الموزعة بشكل موحد ومستقل non–iid. حيث تقوم هذه الطريقة بتقسيم المستخدمين (الذين يملكون بيانات التدريب) إلى مجموعات لديها توزيعات بيانات متشابهة ضمن نفس المجموعة وذلك عن طريق ميزات محددة (بما في ذلك الخسارة التجريبية المحلية عند تطبيق نموذج التعلم الآلي على بيانات المستخدم، أو أوزان النموذج أو تحديثات التدرجات الناتجة)، ومن ثم القيام بعملية التدريب بشكل تعاوني بين المستخدمين ولخادم المركزي.

الدراسات السابقة:

تختلف الدراسات في مجال CFL (Clustered Federated Learning) التعلم الموحد القائم على العنقدة بشكل أساسي في عملية تحديد الزبائن ذوي توزيعات البيانات المتشابهة. حيث اقترح الباحثون في الدراسة [3] تقسيم ثنائي متكرر للزبائن إلى عناقيد عن طريق استغلال تشابه التجيبcosine similarity بين تحديثات النماذج المحلية. لكن هذه الطريقة ليست ذات كفاءة اتصال لأن العنقود يتم تقسيمه فقط بعد أن تتقارب النماذج المحلية للزبائن المرتبطة به.

بينما اقترحت الدراسة [4] استخدام الخسارة التجريبية المحلية للزبائن كمقياس لقياس تشابه الزبائن من حيث توزيع البيانات. إلّا أنّ الطريقة المقترحة تتطلب اتصالات ثابتة بين الزبائن والمخدم لبناء عناقيد مستقرة مما يحتاج كلفة اتصال عالية. بالإضافة إلى أن الأداء يتأثر بشكل كبير بالإعدادات المسبقة لعدد العناقيد الذي من الصعب تحديده دون معرفة توزيعات البيانات المحلية.

بالإضافة إلى وجود عدة دراسات [5,6] تحدد عناقيد الزبائن بالاعتماد على مسافات النموذج المحسوبة من أوزان النموذج. ومع أن هذه الطرق تحتاج لعدد أقل من الاتصالات بين الخادم والزبائن من أجل التجميع إلّا أنّها تحتاج لتحديد عدد العناقيد مسبقاً وهو ما تهدف هذه الدراسة إلى الاستغناء عنه وجعل منهجية التعلم الموحد أكثر ديناميكية ليناسب البيئات العملية الواقعية.

٢- مشكلة البحث:

إن قيود الخصوصية التي يفرضها التعلم الموحد تتمثل بعدم إمكانية الوصول إلى البيانات عند المستخدمين بشكل مباشر مما يؤدي إلى ظهور الأخطاء في اكتشاف توزيعات البيانات وتحديد تشابهها من أجل الوصول إلى مجموعات مستخدمين ذات توزيعات بيانات متشابهة إلى حد كبير. وهو ما يؤدي إلى تدهور أداء نماذج التعلم الآلي المستخدمة وزيادة عدد دورات الاتصال للوصول إلى نموذج موحد كلي ذو أداء مناسب

بالإضافة إلى أن خوارزميات العنقدة المستخدمة من أجل تجميع المستخدمين إلى مجموعات متشابهة بناء على بارامترات معينة ماتزال تعاني من عدم القدرة تحديد الميزات المتطرفة التي تحدد انتماء المستخدم للمجموعة.

٣- أهمية البحث وأهدافه:

نظراً لأهمية التعلم الموحد بصفته آلية تدريب تأخذ بعين الاعتبار مشاكل الخصوصية في حوسبة الحافة والبيئات الموزعة يتم العمل على معالجة التحديات التي تواجه تطبيقه عملياً وبشكل خاص في شبكات حوسبة الحافة وأهمها عدم التجانس الإحصائي للبيانات عند المستخدمين المشاركين في التدريب والتي تحد من أداء النماذج وتزيد حمل الاتصال بين الخادم المركزي والمستخدمين.

يقوم هذا البحث بمعالجة مشكلة عدم التجانس الإحصائي عن طريق عنقدة المستخدمين بناء على تشابه بياناتهم بشكل ديناميكي مع أقل عدد بارامترات مطلوب من أجل تحديد العناقيد. وذلك مع دراسة عدم التجانس الإحصائي بشكل أكثر واقعية وهو ما تفتقر له الدراسات السابقة من حيث محاكاة حالة عدم تماثل توزيعات البيانات non-independent and identically distributed (non-iid).

بالتالي فإن هدف هذا البحث:

- دراسة تأثير عدم تجانس البيانات على أداء نماذج التعلم الموحد.
- اقتراح منهجية قادرة على جعل آلية التدريب أكثر ديناميكية وقدرة على اكتشاف التشابه في التوزيعات الإحصائية.
- الحفاظ على أداء التعلم الموحد القائم على العنقدة مع مراعاة حمل الاتصال.

٤- طرق البحث ومواده:

تمت برمجة محاكاة المنهجية المقترحة باستخدام لغة البرمجة python التي تشتهر في مجال الذكاء الصناعي بسبب عدد مكتباتها وسهولة الاستخدام، وذلك في بيئة العمل google colab التي تتيح استخدام notebook jupyter دون الحاجة إلى تثبيت متطلبات اللغة على الحاسب محلياً وهي أحد الخدمات السحابية التي تقدمها شركة google.

٤-١- حوسبة الحافة (الحوسبة الطرفية):

تشير الحوسبة الطرفية إلى التقنيات التي تسمح بإجراء الحوسبة على حافة الشبكة. هنا نعرّف الطرفية "Edge" بأنه أي مورد حوسبة وشبكة على طول المسار بين مصادر البيانات ومراكز البيانات السحابية. على سبيل المثال، الهاتف الذكي هو الحافة بين الأشياء الجسدية والسحابة، والبوابة في المنزل الذكي هي الحافة بين الأشياء المنزلية والسحابة.

يعتمد منطق الحوسبة الطرفية على أن الحوسبة يجب أن تتم بالقرب من مصادر البيانات. يمكن لحوسبة الحافة أن تكون قابلة للتبديل مع حوسبة الضباب، لكن حوسبة الحافة تركز أكثر على جانب الأشياء، بينما تركز حوسبة الضباب أكثر على جانب البنية التحتية. يتوقع أن الحوسبة المتطورة يمكن أن يكون لها تأثير كبير على مجتمعنا مثل تأثير الحوسبة السحابية مستقبلاً.[7]

٤-٢- التعلم الموحد:

يتيح التعلم الموحد للزبائن صياغة نموذج عالمي باستخدام البيانات المحلية المتاحة للزبون فقط. لا يتم مشاركة البيانات مع خادم مركزي أو بين الزبائن. تتم مشاركة نموذج أولي مع جميع الزبائن المؤهلين ويقوم الزبائن بتدريب النموذج بناءً على بياناتهم المحلية. يتم إرسال الأوزان الناتجة بعد التدريب إلى خادم مركزي يقوم بتجميعها وتحديث النموذج العالمي. يتم استخدام النموذج الجديد كنموذج أولي لتكرار التدريب التالي.

٤-٢-١. التعلم الموحد القائم على العنقدة:

تقوم بيئة التعلم الموحد على أساس نموذج كلي وحيد للمستخدمين وهو ما يختلف مع آلية التطبيق الفعلي في البيئات العملية. حيث يمكن أن يكون للمستخدمين أهداف تدريب مختلفة ودمج المعرفة لدى الكل في نموذج واحد سوف تؤثر سلباً على الأداء المحلي.

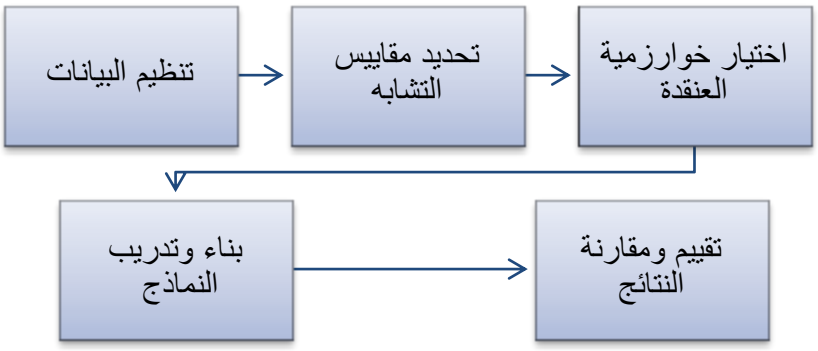
ومنه تم اقتراح التعلم الموحد القائم على التجميع لتجميع المستخدمين إلى عناقيد بناء على التشابه. لكن مع ذلك فإن خوارزميات التعلم الموحد القائم على التجميع الحالية تحتاج إلى افتراض عدد العناقيد مسبقاً بالإضافة إلى أنها غير فعالة كفاية لاكتشاف علاقات التشابه بين بيانات المستخدمين.

٤-٣- تجانس البيانات:

غالباً ما يكون توزيع البيانات على الزبائن (Non-IID) من الناحية العملية، والمعروف أيضاً باسم عدم تجانس البيانات، وهو ما يمثل تحدياً في FL [10]. حيث تعتمد بيانات التدريب في كل زبون بشكل كبير على استخدام أجهزة محلية معينة، وبالتالي، قد يكون توزيع البيانات للعملاء المتصلين مختلفاً تماماً مع بعضهم البعض. تُعرف هذه الظاهرة باسم Non-IID، والتي قد تسبب اختلافاً حاداً في النموذج. [11]

تتعدد حالات Non-IID للبيانات كانحراف السمات أو الفئات أو الكمية. حيث تم اختيار حالتي انحراف الفئات وعدم التوازن الكمي في محاكاة الحالات لهذه الدراسة كونها تحاكي حالات أجهزة حوسبة الحافة التي يتم افتراضها.

٥- مراحل العمل تم العمل على عدة مراحل كما في الشكل(١):



الشكل (١): مراحل العمل

٥-١- مجموعة البيانات:

إن مجموعة البيانات المستخدمة تمثل بيانات المستخدمين الذين يقومون بتدريب نموذج تعلم آلي موحد بشكل مشترك للوصول إلى نموذج واحد قادر على حل مشكلة التصنيف عند أي زبون أو (مستخدم) وذلك دون الحاجة إلى نقل بيانات المستخدمين إلى خادم مركزي ، بالتالي يتم تقييم التدريب باستخدام التعلم الموحد عن طريق قياس أداء نموذج التعلم الآلي المستخدم والذي من الممكن أن يكون أي نموذج تعلم آلي ذو مشكلة مشتركة(تصنيف ، عنقدة ،أو غيرها).

تم الاعتماد على بيانات معيارية من مكتبة tensorflow وهي مجموعات بيانات الغرض منها تقييم أداء آلية العمل الموزعة للتعلم الموحد ومحاكاة بيانات الزبائن المشتركين في التدريب.

مجموعة البيانات CIFAR-10 مكونة من ٦٠٠٠٠ صورة ملونة ٣٢ × ٣٢ في ١٠ فئات، مع ٦٠٠٠ صورة لكل فئة تصنيف، يوجد ٥٠٠٠٠ صورة تدريب و١٠٠٠٠ صورة اختبار.

تنقسم مجموعة البيانات إلى خمس مجموعات تدريب ودفعة اختبار واحدة، كل منها تحتوي على ١٠٠٠٠ صورة. تحتوي مجموعة الاختبار بالضبط على ١٠٠٠ صورة تم اختيارها عشوائياً من كل فئة. تحتوي مجموعات التدريب على الصور المتبقية بترتيب عشوائي، ولكن قد تحتوي بعض مجموعات التدريب على صور من أحد الفئات أكثر من الأخرى. فيما بينها، تحتوي مجموعات التدريب بالضبط على ٥٠٠٠ صورة من كل فصل.

يتم معالجة مسبقة للبيانات والتي هي عبارة عن مجموعة صور بالتالي عمليات معالجة الصور اللازمة من أجل جعل الصور مناسبة كدخل لخوارزميات وشبكات الرؤية الحاسوبية حيث تستخلص خصائص الصورة وتتحول إلى بيانات مناسبة للعمليات الحسابية للنموذج الحاسوبي.[15]

انطلاقا من أن البحث يركز على آلية التدريب في بيئات موزعة دون تبادل بيانات المستخدم فإن عمليات المعالجة المسبقة تقتصر على تحويل البيانات إلى صيغة موحدة ومعايرتها تبعاً لهدف التدريب.

٥-٢- توزيع البيانات:

تركز الأبحاث في هذا المجال على تحويل مجموعة البيانات إلى شكل موزع من حيث توزيع البيانات على المستخدمين من أجل محاكاة البيانات في البيئات الفعلية.

لمحاكاة سيناريو التعلم الموحد نقوم بتقسيم جزء التدريب من كل قاعدة بيانات إلى مجموعات بيانات N تقابل عدد الزبائن مع تقسيم جزء الاختبار أيضاً إلى N مجموعة بيانات. ومن ثم نسند لكل مستخدم قاعدة بيانات مؤلفة من مجموعة تدريب ومجموعة اختبار موافقة من المجموعات السابقة.

آلية التقسيم تتم وفقا لاستراتيجيتين تبعا لطرق أخذ العينات وذلك لمقارنة النتائج وتقييم الفعالية حيث الأولى تراعي خصائصIID والثانية non-iid للبيانات المحلية للزبائن كما هو مبين في الجدول (١):

جدول(١): استراتيجيات تقسيم البيانات

IID	أخذ عدد من العينات بشكل عشوائي تبعاً للفئات من قاعدة البيانات الأصلية.	
Non−iid_5	ترتيب قاعدة البيانات تبعا للفئات وتوزيعها على الزبائن المختلفين حيث يختلف انحراف البيانات بالاعتماد على حجم العينات	كل زبون لديه 5 فئات من البيانات كحد أقصى.
Non−iid_1		كل زبون لديه فئة أو فئتين من البيانات كحد أقصى.

٥−٣− تحديد مقاييس تشابه البيانات:

تختلف بارامترات قياس تشابه بيانات الزبائن في التعلم الموحد عن المقاييس التقليدية حيث أنها تخضع لقواعد الخصوصية المعتمدة في هذه البيئة فهذه المقاييس يجب ألّا تعطي معلومات عن بيانات المستخدمين بالدرجة الأولى. حيث اختلفت الطرق بين ما هو معتمد على تقييم الخسارة وهو أبسطها إلى المعتمد على أوزان النموذج أو بارامتراته.

تم في هذا البحث اقتراح طريقة جديدة لقياس التشابه تقوم على حساب (الانحراف المعياري، عدد العينات وقيمة الخسارة) لتمثيل الزبون.

يعطى الانحراف المعياري بالعلاقة:

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \quad (١)$$

حيث:

y_i: قيمة التسمية (label) للعينة i من مجموعة البيانات المستخدمة(صور).

$\bar{y}$: قيمة متوسط مجموعة البيانات.

n: عدد عينات مجموعة البيانات.

٥−٤− اختيار خوارزمية العنقدة:

استخدمت بعض الدراسات السابقة خوارزميات العنقدة من أجل تجميع المستخدمين إلى مجموعات أو عناقيد ذات خصائص متشابهة بشكل عام وذلك بحسب هدف الدراسة ومتطلبات الخوارزمية , لكن مازال مجال عدم التجانس الإحصائي يعاني من قلة في الأبحاث بسبب قيود الخصوصية التي يفرضها التعلم الموحد من أجل المحافظة على بيانات المستخدمين بالإضافة إلى بعض العيوب التي تعاني منها الخوارزميات المستخدمة من حيث حاجتها للقيام بعدة جولات تدريب قبل الوصول إلى نتيجة عنقدة مناسبة وتحديد عدد العناقيد وعدة بارامترات تحتاج لضبط مسبق وهو مالا يتناسب مع البيئات العملية الديناميكية ومنه تم اقتراح استخدام الخوارزمية التالية في هذا البحث لمواجهة هذه التحديات.

٥−٤−١. خوارزميات التجميع(العنقدة):

HDBSCAN هي خوارزمية تجميع تعتمد على الكثافة يمكنها تحديد مجموعات من أشكال وأحجام مختلفة في البيانات عالية الأبعاد. HDBSCAN تعني التجميع المكاني الهرمي القائم على الكثافة للتطبيقات ذات الضجيج، وهو امتداد لخوارزمية DBSCAN.

مثل DBSCAN، يحدد HDBSCAN المجموعات بناءً على كثافة النقاط في البيانات. ومع ذلك، على عكس DBSCAN، لا يتطلب HDBSCAN من المستخدم تحديد عدد المجموعات مسبقًا. بدلاً من ذلك، يستخدم HDBSCAN نهجاً هرمياً لتجميع البيانات، حيث يتم تشكيل المجموعات عن طريق دمج مجموعات أصغر بناءً على كثافتها وقربها.

يعمل HDBSCAN من خلال إنشاء تسلسل هرمي للمجموعات باستخدام بارامتر يسمى الحد الأدنى لحجم الكتلة. يحدد هذا البارامتر الحد الأدنى لعدد النقاط المطلوبة لتشكيل كتلة. ثم يستخدم HDBSCAN نهجاً قائماً على الكثافة لتحديد الكتل ضمن هذا التسلسل الهرمي، حيث يتم تعريف المجموعات على أنّها مكونات متصلة بمناطق عالية الكثافة.

آلية عمل الخوارزمية:

a. تقوم الخوارزمية بإيجاد مصفوفة مسافات البيانات.

b. يتم إبعاد النقاط ذات الكثافة المنخفضة باستخدام مقياس كثافة جديد (mutual reachability distance)

والذي يعطى بالعلاقة :

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\}$$

حيث المسافة المركزية core distance للبارامتر k من أجل النقطة x تعرف بالشكل $core_k(x)$.

c. بناء حد أدنى لل spanning tree.

d. بناء هرمية العناقيد.

e. تجميع شجرة العنقود (تكثيف).

f.قياس استقرار المجموعات في كل عنقود باستخدام بارامتر الاستمرار $\lambda = \frac{1}{distance}$ ومقياس الاستقرار $\sum_{p \epsilon cluster}(\lambda_p - \lambda_{birth})$حيث تعبر القيمتان عن بداية التقسيم ونهاية اختيار العنقود.

تتمثل إحدى مزايا HDBSCAN في قدرتها على التعامل مع البيانات بكثافات متفاوتة، والتي يمكن أن تشكل تحدياً لخوارزميات التجميع الأخرى. يمكن لـ HDBSCAN أيضاً معالجة البيانات المزعجة ويمكنه تحديد النقاط التي لا تنتمي إلى أي مجموعة، والتي يشار إليها باسم القيم المتطرفة.[16]

بشكل عام، تعد HDBSCAN خوارزمية تجميع قوية يمكن أن تكون مفيدة في مجموعة متنوعة من التطبيقات. إنّ قدرتها على تحديد مجموعات الأشكال والأحجام المتغيرة والتعامل مع البيانات ذات الكثافة المتفاوتة تجعلها أداة قيمة لتحليل البيانات الاستكشافية والتعلم الآلي.

٥-٥- بناء وتدريب النماذج:

يعتمد بناء النموذج على المهمة التي نحتاج للقيام بها باستخدام التعلم الموحد، بالتالي فإن الشبكات العصبية التلافيفية Convolutional Neural Network تعتبر النموذج الأفضل في مجال الدراسة الحالية حيث تقوم باستخدام مجموعة بيانات الصور لدراسة الأداء أي إن مهمة نموذج التعلم الآلي المدرب باستخدام التعلم الموحد هي تصنيف الصور.

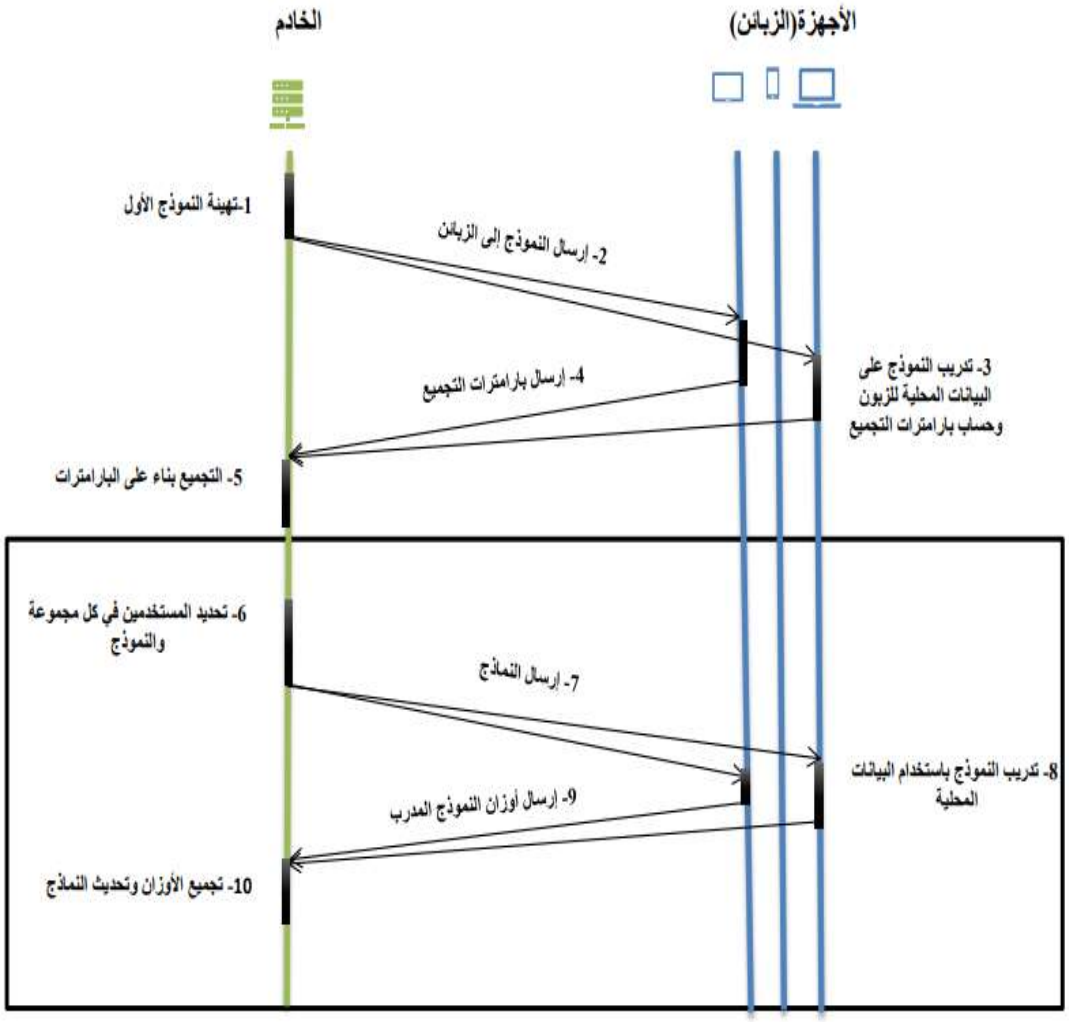
تتكون شبكة CNN عادةً من ثلاث طبقات: طبقة تلافيفية، وطبقة تجميع، وطبقة متصلة بالكامل (a convolutional layer, a pooling layer, and a fully connected layer).

تم اعتماد نموذج الشبكة العصبية التلافيفية المستخدم في الدراسة [٨] المؤلف من ثلاث تكرارات من (طبقتي conv2D مع تابع تفعيل relu متبوعة ب MaxPool2D) متبوعة بطبقة Flatten وطبقتي Dense مع تابع تفعيل relu و softmax على التوالي.

تم ضبط قيم البارامترات تبعاً للدراسة السابقة [12] كالتالي: learning_rate=0.001 لحالة البيانات الأولى و0.00001 للحالة الثانية و batch_size= 10 مع optimizer=' adam'.

٥-٦- الآلية المقترحة:

يبين الشكل (٢) آلية التدريب المتبعة في هذا البحث حيث أن كل مستخدم مشارك في التدريب يطلق عليه اسم (زبون): ١. يقوم الخادم المركزي بتهيئة نموذج تعلم آلي محدد مناسب للمشكلة التي يجب حلها عند المستخدمين، ٢. يتم إرسال النموذج إلى مجموعة الزبائن المحددين من قبل الخادم، ٣. يتم تدريب النموذج على بيانات الزبون المحلية، ٤. يرسل الزبون نتائج التدريب المحلية وهي قيمة الخسارة بالإضافة إلى قيمة الانحراف المعياري وعدد عينات التدريب المحلية كقيم إحصائية تم اقتراح استخدامها في هذا البحث، ٥. تطبيق خوارزمية العنقدة على القيم السابقة ،٦. وتحديد مجموعات الزبائن وأوزان النموذج الموافق لكل مجموعة زبائن، ٧. إرسال النموذج إلى الزبائن، ٨. تدريب النموذج باستخدام البيانات المحلية، ٩. إرسال أوزان النموذج المدرب عند كل زبون ،١٠. تجميع أوزان النموذج لكل مجموعة زبائن وتحديث النموذج:



الشكل(2): آلية التدريب

٥-٧- تقييم ومقارنة النتائج:

تعتمد الأبحاث الموجودة في مجال هذه الدراسة على تقييم الفعالية بناء على قيمة دقة النماذج النهائية وحمل الاتصال، لكن تظهر الحاجة إلى استخدام مقاييس تقييم أخرى اعتماداً على حالة الدراسة وطبيعة البيانات. بالتالي تم التركيز على حساب قيم الدقة لكل الحالات مع عدد دورات الاتصال كمقياس للمقارنة مع الدراسة السابقة، بالإضافة إلى استخدام المقاييس التالية المناسبة للحالة المدروسة حيث أن الدقة في حالة الفئات غير المتوازنة ممكن أن تكون منحازة للفئات الأكبر:

١- الدقةAccuracy : وتمثل عدد الحالات المصنفة بشكل صحيح بالنسبة لجميع الحالات وذلك وفق:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ all\ predictions} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

٢- الضبط Precision: ويمثل عدد الحالات الإيجابية المصنفة بشكل صحيح بالنسبة لجميع الحالات الإيجابية وفق:

$$Precision = \frac{Number\ of\ true\ positives}{Number\ of\ all\ positive\ predictions} = \frac{TP}{TP+FP} \quad (3)$$

٣- الاستدعاء Recall: عدد الحالات الإيجابية المصنفة بشكل صحيح بالنسبة لجميع الحالات الإيجابية وفق:

$$Recall = \frac{Number\ of\ true\ positives}{Number\ of\ all\ positives} = \frac{TP}{TP+FN} \quad (4)$$

حيث توفر الضبط والاستدعاء رؤية أكثر تفصيلاً لأداء النموذج على الأمثلة الإيجابية والسلبية. يقيس الضبط نسبة الإيجابيات الحقيقية بين الأمثلة التي تم التنبؤ بها على أنها إيجابية، بينما يقيس الاستدعاء نسبة الإيجابيات الحقيقية بين جميع الأمثلة الإيجابية.

٥-٧-١. النتائج والمناقشة:

تم استخدام قاعدة البياناتCIFAR 10 المستخدمة مع الدراسة السابقة[12] ,والتي تم توصيفها في الفقرة (٥-١) واختيار قيم البارامترات في هذه الدراسة لغاية المقارنة معها والتي اعتمدت على أوزان النموذج لتمثيل الزبون وهو ما يعتبر عبء حسابي تم العمل على التخلص منه في هذه الدراسة باستخدام قيم تمثيل جديدة قادرة على تمثيل الزبون بشكل أفضل من حيث البيانات أيضاً.

يبين الجدول (٢) مقارنة مع الدراسة السابقة والتي تعتبر الدراسة الوحيدة ذات المفاهيم المشتركة مع دراستنا حيث أن قيم النواحي المذكورة تعبر عن مشكلة وهدف البحث:

الجدول(٢):نواحي التحسين المقدمة في الدراسة

	الدراسة الحالية	الدراسة السابقة
أنواع توزع البيانات	٣	٢
عدد بارامترات خوارزمية العنقدة	١	٢
عدد البارامترات المتبادلة لقياس التشابه	٣	عدد أوزان النموذج

تم الحصول على النتائج مع الإعدادات التالية:

١- توزيع بيانات تبعا للجدول (١).

٢- ١٠٠ مستخدم.

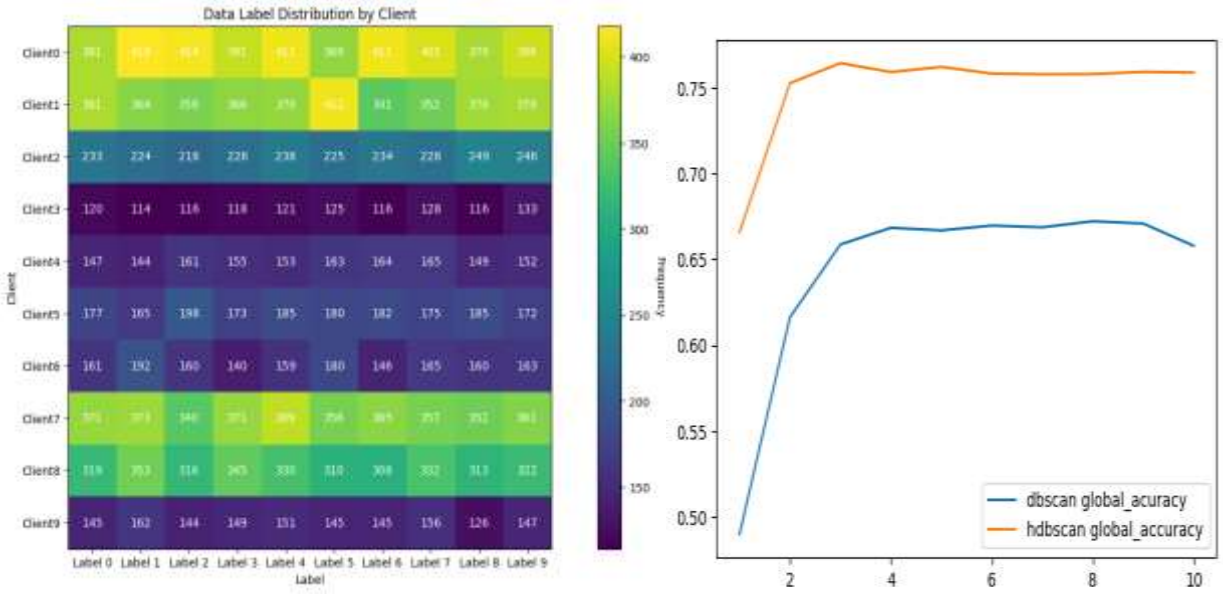
٣- نسبة توافر client_ratio=0.1 أي ١٠ زبائن مشاركين بكل دورة اتصال.

كما تم ضبط توزيع البيانات بحيث تشكل ٣ عناقيد لغاية المقارنة فقط مع الدراسة السابقة مع الحفاظ على عشوائية عملية أخذ العينات من مجموعة البيانات الأساسية تماشياً مع محاكاة البيانات الواقعية لبيانات المستخدمين(الزبائن).

٥-٧-١-١- الحالة الأولىIID:

تمثل الحالة الأولى تدريب النموذج مع بيانات متجانسة إلى حد ما من حيث الفئات لكنها غير متوازنة من حيث عدد العينات عند كل زبون. تمت المحاكاة ل١٠ دورات اتصال مع توزيع بيانات مبين بالشكل(٣) وكانت النتائج كالتالي :

الشكل(٣):مخطط يوضح دقة النموذج خلال ١٠ دورات اتصال مع خريطة حرارية توضح توزع فئات البيانات



من الشكل(٣) نلاحظ تفوق نتائج تدريب البيانات عند المنهجية المقترحة مع خوارزمية HDBSCAN من حيث الدقة النهائية على نتائج التدريب مع خوارزمية DBSCAN وهو ما يفسر بقدرة الخوارزمية الأولى على فصل العناقيد وتحديد البيانات الشاذة بشكل أفضل مما يؤثر على النتائج النهائية التي تتأثر بتدريب عناقيد متشابهة من بيانات الزبائن كنتيجة لمرحلة العنقدة.

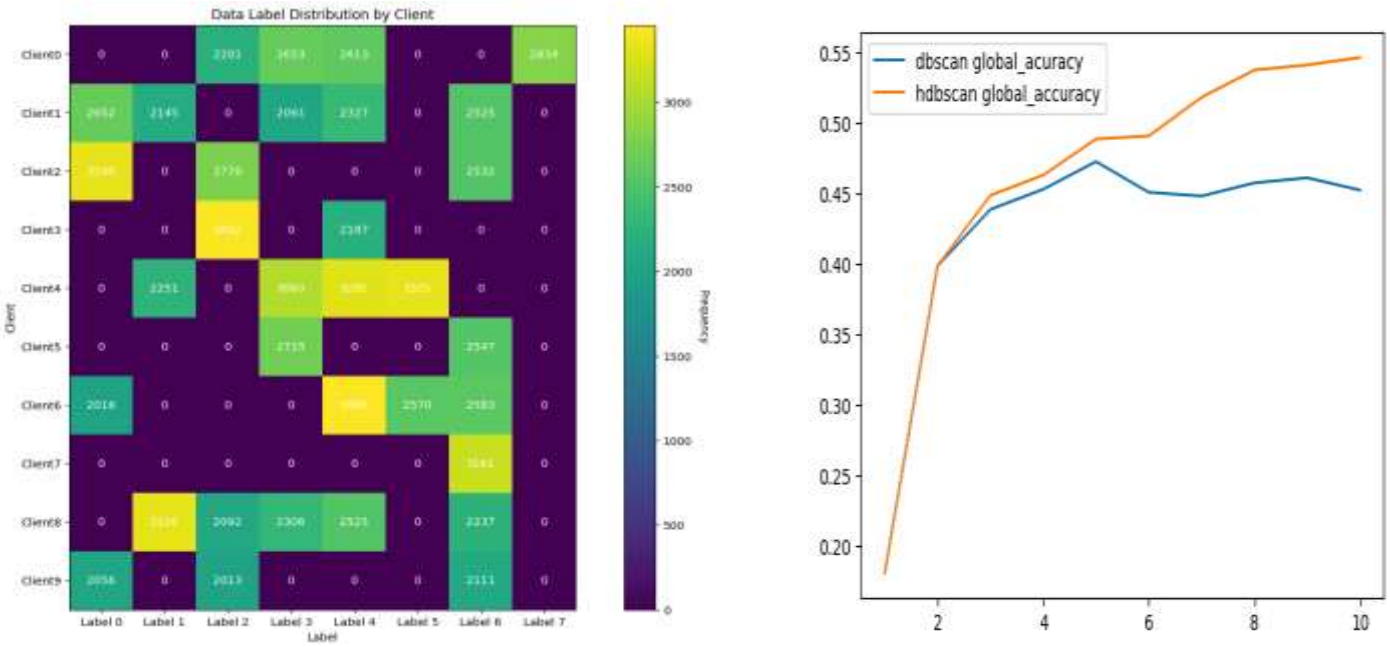
الجدول(٣): مقارنة أداء النموذج المقترح مع خوارزميتي عنقدة بالنسبة للدراسة السابقة

عدد دورات الاتصال	Recall	Precision	Global Accuracy	
١٠	75 %	77%	75.8%	hdbscan
١٠	64.2%	68.1%	٦٥.٧%	dbscan
١٠٠	–	–	39.7%	الدراسة السابقة

توضح النتائج الواردة في الجدول (٣) أفضلية المنهجية المقترحة من عدة نواحي أولها تحسن الدقة بشكل ملحوظ جداً في هذه الحالة وعند عدد دورات اتصال أقل بفارق كبير ٩٠ دورة. بالإضافة إلى قيم أفضل لكل من الاستدعاء والضبط. تعزى هذه الفروق إلى استخدام بارامترات قياس تشابه مناسبة تجعل خطوة تحديد التشابه خطوة جوهرية للوصول إلى نتائج أفضل حتى في حالة البيانات شبه المتجانسة.

٥-٧-١-٢- الحالة الثانية (Non-iid_5):

تمثل الحالة الثانية تدريب النموذج مع بيانات غير متجانسة بنسبة ٠.٥ أو ٥ فئات كحد أقصى عند كل زبون مع عدد عينات متفاوت عند كل زبون. تمت المحاكاة ل١٠ دورات اتصال مع توزيع بيانات مبين بالشكل (٤) وكانت النتائج كالتالي:



الشكل(٤): مخطط يوضح دقة النموذج خلال ١٠ دورات اتصال مع خريطة حرارية توضح توزع فئات البيانات

نلاحظ تفوق خوارزمية HDBSCAN أيضاً من حيث دقة النموذج الكلي عند زيادة عدد مرات التدريب حيث تم الاختبار لمدة ١٠دورات اتصال.

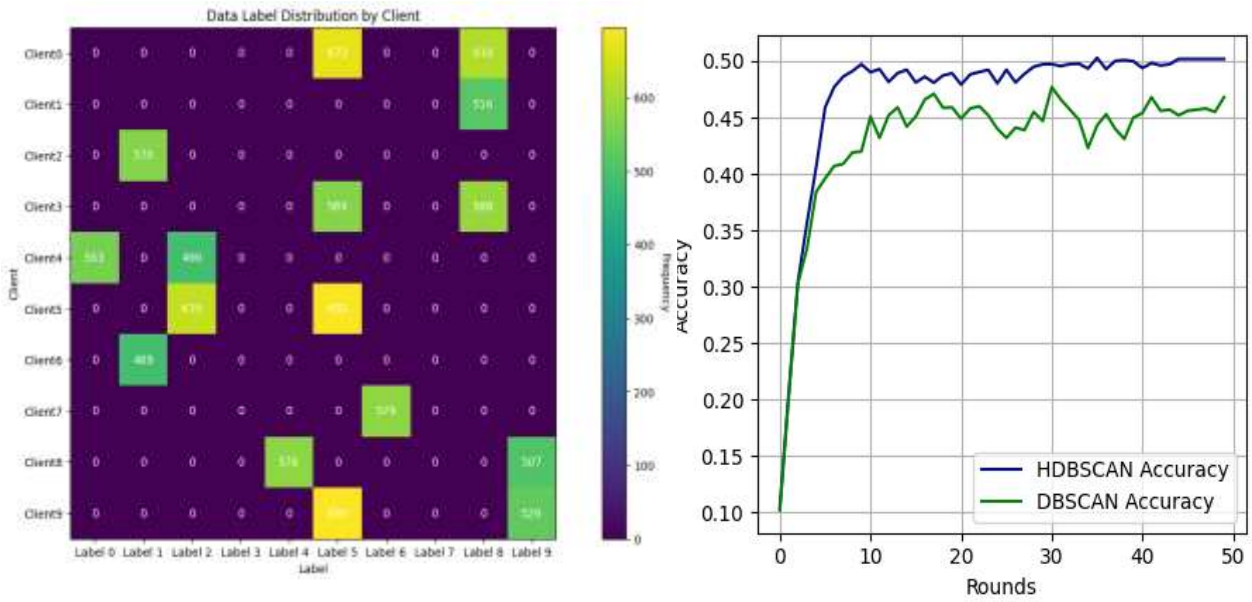
الجدول (٤) مقارنة أداء النموذج المقترح مع خوارزميتي عنقدة بالنسبة للدراسة السابقة

Recall	Precision	Global Accuracy	خوارزمية العنقدة المستخدمة
43%	48%	50%	hdbscan
42.5%	49.4%	48.5%	dbscan

يبين الجدول (٤) أفضلية المنهجية المقترحة من حيث الدقة الكلية بالإضافة إلى الاستدعاء والضبط. كما يبين قيمة ضبط أفضل لDBSCAN وهو ما يمكن أن يعود إلى عشوائية أخذ عينات بيانات كل مستخدم.

٥-٧-١-٣- الحالة الثالثة (Non-iid_1):

تمثل الحالة الثالثة تدريب النموذج مع بيانات غير متجانسة بنسبة ٠.١ أو فئتين كحد أقصى عند كل زبون مع عدد عينات متفاوت. تمت المحاكاة ل ٥٠ دورة اتصال مع توزيع بيانات مبين بالشكل (٥).



الشكل (٥): مخطط يبين قيم دقة النموذج خلال ٥٠ دورة اتصال مع توزيع البيانات المقابل

نلاحظ من المخطط السابق فعالية خوارزمية العنقدة HDBSCAN المستخدمة في المنهجية المقترحة من حيث الدقة النهائية مقارنة مع DBSCAN في حال وجود بيانات ذات درجة كبيرة من عدم التجانس، وذلك خلال ٥٠ دورة اتصال لغرض المقارنة وتثبيت النتائج على مدى طويل من دورات التدريب للتأكد من القيم.

في الدراسة السابقة تم قياس الدقة لكل عنقود على حدى لكن بسبب عشوائية البيانات لا يمكننا مقارنة النتائج بهذه الطريقة لذلك سنقوم باستخدام الطريقة المقترحة في الدراسة [9] والتي تقوم على حساب متوسط الدقة للعناقيد ومقارنها مع نتائج الطريقة المقترحة بنفس الطريقة.

الجدول (٥): مقارنة نتائج التنفيذ مع الدراسة السابقة للحالة Non-iid_1 بعد ٥٠ دورة اتصال

خوارزمية العنقدة المستخدمة	بارامترات قياس التشابه	Global Accuracy	Average Test Accuracy	Precision	Recall
hdbscan	Standard deviation, loss value, sample size	50%	23.5%	56%	46%
dbscan		46.5%	21.6%	54.2%	42%
الدراسة السابقة	Model parameters	-	21.3%	-	-

توضح النتائج الواردة في الجدول (٥) أفضلية المنهجية المقترحة بالنسبة للدراسة السابقة من ناحية تحسن متوسط الدقة بشكل ملحوظ وعند عدد دورات اتصال واحد. بالإضافة إلى قيم أفضل لكل من الاستدعاء والضبط عند استخدام HDBSCAN للمنهجية المقترحة. تؤكد أفضلية هذه القيم على فعالية المنهجية في حالات عدم تجانس البيانات من حيث القيم الأفضل للدقة ومن خلال تخفيف عبء الاتصال في جميع الحالات المدروسة.

من النتائج السابقة نلاحظ تقارب نتائج استخدام خوارزميتي العنقدة DBSCANوHDBSCAN بشكل ملحوظ بالنسبة للدراسة السابقة لكن وتبعاً لهدف الدراسة نستنتج أن HDBSCAN لها أفضلية من ناحية أخرى وهي عدد البارامترات التي تحتاج تحديدها قبل بدء التدريب وهو ما يحقق أهداف الدراسة بجعل العملية أكثر ديناميكية وواقعية تماشياً مع تغير البيانات في البيئات العملية.

٦- الاستنتاجات والتوصيات:

قامت الدراسة بمحاكاة عدم تجانس بيانات الزبائن المشاركين في تدريب نماذج التعلم الآلي في التعلم الموحد ودراسة تأثيره على الأداء في بيئة حوسبة الحافة من منظور محدد ضمن حدود دراسة معينة وقد حققت الطريقة المقترحة أداء أفضل في الحالات الثلاث بما فيها حالة البيانات المتجانسة.

حيث أظهرت أن اختيار بارامترات تمثيل الزبون مع خوارزمية عنقدة مناسبة له دور كبير في الوصول لنتائج مهمة مع تخفيض كبير في حمل الاتصال حيث أن التمثيل الصحيح للزبون يعطي صورة أوضح لخوارزمية العنقدة التي بدورها تعطي نتائج أفضل مع الإشارة إلى أن الطريقة المقترحة تسعى للوصول إلى الديناميكية في جميع مراحل العمل تناسباً مع البيئات العملية.

انطلاقاً من السعي المستمر للوصول إلى حلول أمثلية تحاكي المشكلة نقترح كتوصيات مستقبلية:

- العمل على تمثيل الزبون باستخدام بارامترات أخرى تتعلق بمواصفات الأجهزة الطرفية أو الزبائن كمستوى الطاقة والقدرة الحسابية وهو ما يحدد مواصفات أجهزة بيئة حوسبة الحافة التي تعمل تحت قيود محدودية الحوسبة والطاقة وتوافرية الشبكة.
- العمل على استخدام تقنيات تشفير بيانات التشابه والأوزان المتبادلة عبر الشبكة وذلك لتحقيق الغاية من التعلم الموحد بالشكل الأمثل.
- تمثل هذه الدراسة نواة يمكن البناء عليها من أجل اختبار المنهجية مع أنواع بيانات أخرى حيث عالجت الدراسة مجموعة بيانات صور.

٧- المراجع:

[1] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., & He, B. (2019). *A Survey on Federated Learning Systems: Vision. H*ype and Reality for Data Privacy and Protection.

[2] Wang, Y., Wolfrath, J., Sreekumar, N., Kumar, D., & Chandra, A. (2021, April). *Accelerated training via device similarity in federated learning.* In Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking (pp. 31-36).

[3] F. Sattler, K.-R. Muller, and W. Samek. Clustered feder- ¨*ated learning: model-agnostic distributed multitask optimization under privacy constraints*. IEEE Transactions on Neural Networks and Learning Systems, 32(8):3710– 3722, 2021.

[4] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. *An efficient framework for clustered federated learning*. Advances in Neural Information Processing Systems, 33:19586–33:19586–19597, 2020.

[5] C. Briggs, Z. Fan, and P. Andras. *Federated learning with hierarchical clustering of local updates to improve training on non-IID data.* In 2020 International Joint Conference on Neural Networks, pages 1–9, 2020

[6] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing. ClusterFL: *a similarity-aware federated learning system for human activity recognition. In Proceedings of the Annual International Conference on Mobile Systems, Applications, and Servic*es, pages 54–66, 2021.

[7] Shi, Weisong, et al. "Edge computing: Vision and challenges." *IEEE internet of things journal* 3.5 (2016): 637-646.

[]McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.* 2.11 (2017): 205.

[9] Raveen Bandara Harasgama, Pulasthi. "*Cluster selection for Clustered Federated Learning using Min-wise Independent Permutations and Word Embeddings*." (2022).

[10] Yao, D., Pan, W., Dai, Y., Wan, Y., Ding, X., Jin, H., ... & Sun, L. (2021*). Local-global knowledge distillation in heterogeneous federated learning with non-iid data.* arXiv preprint arXiv:2107.00051.

[11] Zhu, H., Xu, J., Liu, S., & Jin, Y. (2021). Federated learning on non-IID data: A survey. *Neurocomputing*, *465*, 371-390.

[12] Lu, C., Deng, S., Wu, Y., Zhou, H., & Ma, W. (2022). Federated Learning Based on OPTICS Clustering Optimization. *Discrete Dynamics in Nature and Society*, *2022*.

[13] Ibrahim, M. M., & Rizvi, S. A. M. *A Metasystem Base Representation of the Standard System Data Dictionary. e*xchange, 7, 8

[14] Mr. Maher Ibrahim & Dr. S.A.M. Rizvi, *A hybrid Approach for Better Utilization of Metadata Schema Registries, f*irst national conference on next generation computing & information systems, CD proceedings under knowledge and Data Engineering section, S. No: 9, 12-13 May 2007, Model Institute of Engineering and Technology (MIET), Jammu, Kotbhalwal, India.

[15] Salman, Jaafar; Saad, Ghada; Suliman, Marie. 2021, *Using deep learning algorithms and computer vision in detecting human brain tumor*, Tartous University Journal, Vol. 5, No. 10.

[16] How HDBSCAN *Works. Accessed: Mar. 21*, 2023. [Online]. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.