Politecnico di Torino

Politecnico di Torino

Laurea Magistrale in Ingegneria Biomedica A. A. 2021/2022

Sessione di Laurea Dicembre 2022

Machine learning per la diagnosi della malformazione di Arnold-Chiari: rilevanza dei sintomi e dei parametri morfologici

Relatore: Mesin Luca

Candidato: Briccarello Viola

Abstract

La sindrome di Arnold-Chiari è una malformazione della fossa cranica posteriore e come molte altre malattie rare è ancora oggi oggetto di ricerca.

Il metodo di diagnosi attuale consiste nella visualizzazione di immagini di risonanza magnetica sagittale dei pazienti e nella misurazione manuale della discesa della fossa tonsillare da parte dei neurochirurghi: questo approccio, però, è soggettivo ed esposto ad un certo grado di incertezza; infatti, basandosi soltanto su tale valore si commettono spesso errori di valutazione, svolgendo così interventi chirurgici non adeguati oppure non necessari.

Per questo motivo, l'utilizzo di metodi automatici di diagnostica potrebbe aumentare l'accuratezza e l'efficacia della diagnosi e della cura.

Questa tesi ha lo scopo quindi di indagare la rilevanza di determinate informazioni nel trattamento della CM I tramite il confronto di differenti metodi di classificazione a cui verranno applicati dati di input differenti:

- I primi includeranno soltanto i parametri morfologici;
- I secondi includeranno soltanto la sintomatologia preoperatoria;
- I terzi includeranno entrambi i fattori.

Le immagini di risonanza magnetica sono sopposte a pre-processing e segmentazione delle aree cerebrali di interesse quali cervello, fossa posteriore e cervelletto.

Successivamente, vengono ricavati i parametri morfologici caratteristici, derivanti dalla geometria della fossa cranica posteriore e dalla misura delle aree delle tre strutture cerebrali di interesse.

I sintomi invece sono estratti direttamente dalle cartelle cliniche dei medesimi pazienti.

L'obiettivo finale del progetto è quello di poter comprendere quali parametri siano più utili per poter fornire una previsione affidabile dell'outcome dell'intervento di decompressione cranio-cervicale e di definire quale sia il percorso di cura più adeguato ad ogni singolo paziente.

Sommario

1. Introduzione

1.1 Definizioni e cenni storici

All'alba della scoperta della Malformazione di Chiari I (CM I), la definizione di questa patologia era poco chiara, in quanto non fu inizialmente stabilita una soglia minima relativa al volume di ernia cerebellare necessario per la diagnosi, ma soltanto la presenza di erniazione tonsillare asimmetrica era ritenuto un parametro sufficiente per descrivere tale anomalia.

In seguito, studi condotti da Abulezz e Barkovich (1885) ipotizzarono che, in riferimento alla popolazione sana, le tonsille cerebellari potessero erniare fino ad un limite pari a 3 mm, sostenendo quindi che il valore minimo di erniazione dovesse essere tale; successivamente, è stato stabilito che il cut-off inferiore per le ernie della CM I fosse pari a 5 mm^2.

Il documento di Consenso internazionale ribadisce che una discesa inferiore ai 3 mm sia da considerarsi come una semplice variante fisiologica e che un'erniazione compresa tra 3-5 mm sia invece da ritenersi borderline, con successiva necessità di follow up radiologico ristretto solamente ai casi sintomatici oppure associati a siringomielia, ad un profilo tonsillare appuntito o ad un affollamento dello spazio subaracnoideo a livello della giunzione cranio-cervicale.

Pertanto, ad oggi la Malformazione di Chiari I è definita come una discesa di una o di entrambe le tonsille cerebellari rispetto alla linea di McRae.

La Sindrome di Chiari invece è da considerarsi come la manifestazione clinica della suddetta malformazione. Nello specifico, secondo il documento di Consenso internazionale, per la definizione di Chiari sintomatica sono necessari almeno 2 dei criteri clinici diagnostici di seguito elencati:

- Cefalea di breve durata (minore di 5 minuti) e provocata o peggiorata da tosse.
- Sintomi e segni del tronco cerebrale, cerebellari e/o disfunzioni del midollo cervicale.
- Sintomi e segni oto-neurologici (come vertigini o perdita di equilibrio).
- Scoliosi (criterio opzionale).

1.2 Classificazioni

A causa della varietà clinica che la malformazione può provocare, nel corso degli anni sono state proposte numerose classificazioni che potessero essere utili nella scelta del diverso tipo di trattamento. Classicamente però, la CM è stata suddivisa in quattro tipologie distinte, basate

sulla presentazione morfologica e sulla gravità dei difetti anatomici visualizzabili all'imaging (o all'autopsia).

La Chiari di tipo I è la forma che più spesso si diagnostica in modo occasionale ed è caratterizzata dalla erniazione di una o di entrambe le tonsille cerebellari, le quali appaiono appuntite e protrudono di almeno 5 mm al di sotto del forame magno. In casi molto rari, quando la Chiari è di tipo I, è possibile che si assista ad una risoluzione spontanea dell'anomalia, sia in epoca pediatrica sia in quella adulta.

La malformazione di Chiari di tipo II è invece contraddistinta dalla migrazione caudale del tronco encefalico, del cervelletto e del quarto ventricolo attraverso il forame magno unitamente allo spostamento del midollo spinale cervicale verso il basso. Questa categoria è sempre associata alla presenza di spina bifida aperta o cistica e spesso anche a siringomielia ed idrocefalo.

La Chiari di tipo III è definita dall'erniazione del cervelletto, con l'eventuale inclusione del tronco cerebrale e dalla presenza contemporanea di un meningoencefalocele occipitale basso o cervicale alto.

La malformazione di Chiari di tipo IV è in ultimo catalogata come una grave ipoplasia cerebellare correlata ad un encefalocele occipitale: questa definizione ha però iniziato ad essere considerata obsoleta.

Col tempo sono stati riconosciuti altri due sottoinsiemi di CM, con prognosi differenti. La Chiari di tipo 0 è contraddistinta dalla presenza di siringomielia con minima (< 3 mm) oppure assente erniazione del rombencefalo. Questa tipologia risulta essere adeguata e ben responsiva all'intervento di decompressione della fossa cranica posteriore.

La Malformazione di Chiari di tipo 1.5 è invece una forma più complicata di CM I, caratterizzata dalla formazione di un'ernia tonsillare con l'allungamento e la discesa del tronco cerebrale e dell'obex, situati al di sotto della linea di McRae. Nonostante CM I e CM 1.5 presentino somiglianze morfologiche e anatomiche, è necessario fare una distinzione radiologica accurata perché l'outcome post-operatorio è spesso del tutto diverso. Secondo Tubbs et al., infatti, i pazienti con CM 1.5 hanno maggiore probabilità di andare incontro ad un fallimento dell'operazione di decompressione della fossa cranica posteriore e, solitamente, si assiste alla persistenza di siringomielia.

A giugno 2021 poi, è stato pubblicato sulla rivista Neurological Sciences il "documento di consenso internazionale sulla diagnosi ed il trattamento di Chiari e Siringomielia negli adulti",

in cui è stata fornita una nuova classificazione aggiornata, ritenuta più adeguata ed efficace a descrivere le differenti tipologie di CM e che attualmente è la guida prediletta per la selezione del trattamento chirurgico più adeguato. In questa differenziazione sono state considerate solamente le CM1 e CM2 con riferimento alla descrizione originale di Hans Chiari mentre non sono state incluse le tipologie III e IV perché ritenute anomalie embrionarie estremamente severe, ultra-rare e non correlate alle prime due forme.

1.3 Aspetti clinici della CM I

Le manifestazioni cliniche derivanti dalla malformazione di Chiari di tipo 1 sono piuttosto variegate e riflettono almeno in parte le alterazioni morfologiche riscontrabili dall'imaging: possono infatti essere correlate al restringimento del forame magno dovuto all'ectopia tonsillare e all'ostruzione della circolazione del liquido spinale. Il sintomo decisamente più comune è la cefalea che, in questo caso, ha caratteristiche specifiche e peculiari: infatti viene spesso descritta come un dolore sordo, pulsante, esplosivo ed a volte addirittura lancinante, di durata variabile e che si localizza perlopiù in specifiche aree del cranio. In alcuni casi invece, il dolore si localizza sulla spalla, sulla schiena e sugli arti superiori e inferiori, limitandone così i movimenti. Inoltre, la cefalea è aggravata da attività capaci di aumentare la pressione intracranica, come ad esempio la flessione e l'estensione del collo, la tosse e gli starnuti, il pianto o la risata, l'esercizio fisico. Studi condotti su questi pazienti hanno dimostrato una differenza di pressione tra i ventricoli e lo spazio subaracnoideo lombare: il dolore causato dalla tosse potrebbe, quindi, essere provocato dalla compressione di strutture anatomiche nocicettive nello spazio aracnoideo o sui vasi sanguigni che circondano l'ernia tonsillare.

Questa teoria è inoltre supportata dal fatto che, dopo l'applicazione dell'intervento chirurgico di decompressione, spesso scompaiano la differenza di pressione craniospinale ed anche la cefalea da tosse.

Altri sintomi frequenti, legati ad una tensione delle strutture situate nella fossa cranica posteriore, sono le vertigini, la nausea, gli acufeni, i disturbi vestibolari, il dolore al collo e l'intorpidimento nucale. Le manifestazioni cliniche che possono invece essere associate alla compressione dei nervi cranici caudali comprendono le palpitazioni, la raucedine e la disfagia. Tra i sintomi comuni attribuiti invece ai pazienti pediatrici con CM I si annoverano la cefalea, le difficoltà respiratorie e le alterazioni sensoriali, tra cui intorpidimento (anche facciale).

Meno frequentemente invece si osservano goffaggine, aritmie, disartria, disfunzione del nervo cranico inferiore e l'incontinenza urinaria. Queste presentazioni possono ritardare la diagnosi o confonderla con altre patologie come la paralisi cerebrale, in particolare nei neonati.

In letteratura è stato dimostrato che la deformità spinale sia un'altra importante manifestazione clinica della malformazione di Chiari I e della siringomielia.

Per quanto riguarda le altre forme di CM, invece, esse hanno caratterizzazioni cliniche simili a quelle presentate nel caso della CM I, che comprendono quindi la debolezza degli arti, la scoliosi e le parestesie; inoltre, sono tipicamente presenti cefalee posteriori di breve durata.

Per quanto concerne la CM 1.5 sono stati segnalati più frequentemente sintomi atipici come la cefalea in assenza di manovre responsabili di un aumento della pressione intracranica, mancanza di respiro, dolore alla mascella, riflesso del vomito assente, letargia e disturbi dell'equilibrio con cadute.

La stretta correlazione tra la Malformazione di Chiari di tipo I ed i disturbi del sonno è ormai nota da tempo. Già List nel 1941 pubblicò un numero di articoli inerenti alle malformazioni della transizione cranio-cervicale (MCCT), analizzandone le manifestazioni cliniche, e da questi si comprende come, in una buona percentuale di pazienti, con il tempo si assiste allo sviluppo di disturbi respiratori durante il sonno che possono arrivare a peggiorare la qualità di vita dei soggetti fino a dover ricorrere all'intervento chirurgico. Queste manifestazioni solitamente si interrompono grazie ad un brusco risveglio del paziente ma sono comunque accompagnati da sintomi diurni, come l'eccessiva sonnolenza, l'astenia e il calo dell'attenzione che possono andare lo stesso a compromettere la qualità di vita dei pazienti. Pertanto, si può affermare che, nell'ambito delle Malformazioni di Chiari, il disturbo respiratorio più frequentemente riportato sia l'apnea notturna, nonostante questo sia un sintomo raro e, nella maggior parte dei casi, questa disfunzione migliora in seguito alla decompressione chirurgica della fossa cranica posteriore. Tuttavia, ci sono casi di manifestazioni di apnea post- operatoria in pazienti pediatrici ed una delle complicanze più temute in seguito ad un intervento di questo tipo è proprio la depressione respiratoria a seguito di un intervento chirurgico decompressivo entro i primi cinque giorni dall'operazione. La causa di alcuni decessi post- operatori può essere proprio ricondotta all'apnea, la quale può comparire in seguito alla ricomparsa di erniazione delle tonsille cerebellari che costringe ad effettuare un secondo intervento.

Una revisione del 2020 ha inoltre valutato la frequenza della comparsa di sintomi associati alla Malformazione di Chiari di tipo I, suddividendoli in tre categorie: sintomi relativi all'ostruzione

del liquido cerebrospinale, sintomi dovuti alla compressione della giunzione craniocervicale e sintomi riconducibili alla disfunzione spinale e siringomielia. Da questo studio è infine emerso come la cefalea si confermi essere la manifestazione sintomatica più frequente e, in base alle sue caratteristiche, si può inoltre fare una suddivisione delle differenti tipologie di cefalee correlate a diversi fattori, come ad esempio a tosse, ad emicrania oppure connessa ad entrambe ed anche a dolore irradiato a collo e spalle. Gli altri sintomi considerati essere i più frequenti in letteratura sono le parestesie, i disturbi vestibolari, la disfagia, l'apnea e la nausea; raramente sono state poi osservate cadute. In conclusione, alcuni pazienti con CM-I asintomatica sviluppano sintomi successivi ad un trauma e ci sono anche alcune segnalazioni legate a decessi avvenuti durante lo svolgimento di attività fisica e sportiva.

1.4 Siringomielia

La siringomielia è una condizione neurologica provocata dalla formazione di una cavità piena di liquido all'interno del parenchima del midollo spinale o del canale centrale, dovuta ad un disturbo della circolazione del liquido cefalorachidiano. Tipicamente è associata alla Malformazione di Chiari di tipo 1 ma è possibile osservarne l'insorgenza in seguito a tumori del midollo spinale, traumi e aracnoiditi adesive post-traumatiche o infettive. La siringomielia può inoltre coinvolgere qualsiasi tratto del midollo ed è possibile analizzare le sue caratteristiche soprattutto attraverso le scansioni di risonanza magnetica pesate T2.

Sebbene la siringomielia si manifesti prevalentemente attraverso sintomi sensoriali, come dolore ed insensibilità alla temperatura, nella maggior parte dei casi però, si tratta di una conseguenza dell'applicazione della risonanza magnetica, sempre più frequentemente utilizzata per la valutazione di routine della lombalgia e del dolore cervicale. La maggior parte dei pazienti che presenta tale anomalia, ha un'età compresa tra i 20 ed i 50 anni ed è stata inoltre suggerita una componente genetica della trasmissione.

La storia clinica dei pazienti affetti da siringomielia unitamente a CM I è variabile ed imprevedibile, caratterizzata da periodi di stabilità e di progressione, in cui il decorso può evolvere nell'arco di mesi o anni fino però ad arrivare ad un rapido deterioramento del quadro clinico.

Anche per la siringomielia esiste una classificazione aggiornata al 2021, presente nel documento di consenso internazionale sulla diagnosi ed il trattamento di Chiari e Siringomielia negli adulti e anche nel caso di questa patologia vengono presentati differenti tipi:

- Tipo I: siringomielia con ostruzione a livello del forame magno e dilatazione centrale del canale spinale. Viene sotto-classificata in Tipo IA quando è associata a CM1 e in Tipo I B quando correlata ad altre lesioni ostruenti il forame magno;
- Tipo II: siringomielia senza ostruzioni a livello del forame magno, o idiopatica;
- Tipo III: siringomielia secondaria ad altre affezioni del midollo spinale;
- Tipo IV: idromielia pura, cioè una dilatazione del canale centrale del midollo spinale con sottile cavitazione intramidollare e spesso localizzata in un piccolo segmento.

Poiché la siringomielia è comunemente associata alla CM I, è utile distinguere le caratteristiche cliniche correlate direttamente alla malformazione da quelle invece imputabili alla presenza di una siringa. Nel primo caso i sintomi comuni sono:

- Cefalea tussiva, in sede sub-occipitale, a insorgenza improvvisa, di tipo compressivo e con qualità martellante quando grave ma per il resto non palpitante. Si irradia al vertice e dietro gli occhi e inferiormente al collo e alle spalle, è di breve durata e viene esacerbata dallo sforzo fisico, manovre di Valsalva, flessioni del capo e improvvisi cambiamenti di postura.
- Raucedine, disfagia e tosse accompagnata da deglutizione.
- Disturbi visivi.
- Sintomi oto-neurologici quali vertigini, acufeni, diminuzione dell'udito.
- Disturbi cerebellari come tremori, dismetria, atassia e problemi di equilibrio.
- Sincope.
- Disturbi del sonno come russamento, apnea notturna e palpitazioni.

Le manifestazioni cliniche causate invece solamente dalla siringomielia dipendono dal suo livello anatomico ed includono:

- Sintomi sensoriali: col tempo si assiste anche alla perdita della nocicezione e della termocezione e, a causa dell'assenza della sensazione dolorifica, i pazienti spesso si procurano lesioni e ulcere cutanee croniche sulla mano senza accorgersene.
- Debolezza muscolare con alterata funzione motoria che può progredire fino all'atrofia dei muscoli intrinseci della mano.
- Spasticità agli arti inferiori.
- Perdita del controllo sfinterico.
- Scoliosi progressiva.

- La Sindrome di Horner, che può essere osservata nelle siringhe cervicali/toraciche;
- Dolore neuropatico scarsamente localizzato;
- Ansia, compromissione della memoria, depressione.

La presentazione clinica della siringomielia è, dunque, molto variabile ma nella maggior parte dei casi i pazienti lamentano dolore, debolezza muscolare, atrofia delle mani/delle braccia, insensibilità alla temperatura dell'arto superiore, spasticità o rigidità degli arti inferiori e scoliosi progressiva. Esiste, inoltre, una relazione lineare tra la morfologia della cavità, la durata dei sintomi e la gravità di essi.

1.5 Diagnosi di Malformazione di Chiari: Indagini radiologiche

1.5.1 Imaging prenatale

Per quanto riguarda le procedure di imaging che possono essere impiegate nella diagnosi prenatale di CM I, l'ecografia è attualmente la tecnica di elezione per la valutazione delle anomalie fetali. La risonanza magnetica fetale, infatti, è uno strumento diagnostico di terzo livello che, nonostante abbia una maggiore sensibilità ed affidabilità diagnostica nell'evidenziare la presenza di disrafismo spinale aperto, ha però una capacità limitata di rivelazione delle malformazioni del midollo e della giunzione cranio-cervicale in epoca prenatale. Attraverso l'utilizzo della tecnologia ecografica è possibile invece fare diagnosi precoce e stabilire inoltre se ci sia una qualche indicazione al trattamento chirurgico intrauterino; tale trattamento però viene solitamente riservato a pazienti affetti da Malformazione di Chiari di tipo II dato il fatto che sia stato dimostrato che una riparazione in utero, anche se non ancora totalmente accettata, conduca ad una migliore risoluzione delle ernie tonsillari e ad un maggiore miglioramento della funzionalità motoria rispetto ai casi in cui l'attuazione dell'intervento sia prevista ed eseguita in epoca post-natale.

1.5.2 Imaging postnatale

In epoca post-natale, invece, la valutazione radiografica con proiezione laterale del cranio viene impiegata per indagare l'eventuale presenza di anomalie tipiche della CM 1.5.

Inoltre, le immagini radiografiche della colonna in proiezione antero-posteriore e laterale possono essere necessarie per il calcolo dell'angolo di Cobb coronale e dei punteggi di Risser, parametri necessari ad escludere oppure confermare la presenza contestuale di scoliosi.

La tomografia computerizzata (TC) volumetrica è di solito utilizzata solamente per lo studio delle anomalie ossee della base cranica, della giunzione cranio-cervicale oppure in tutti quei casi in cui le altre procedure risultino controindicate. Infatti, la risonanza magnetica è ad oggi la tecnica di imaging più sensibile per lo studio delle anomalie intracraniche nella Malformazione di Chiari. Normalmente vengono eseguite delle scansioni craniche e dell'intera colonna vertebrale per indagare tutti quei disturbi morfologici che coinvolgono l'intero sistema nervoso centrale (SNC), per valutare la dinamica del flusso del liquido cerebrospinale, per stimare il grado di erniazione tonsillare, per la ricerca di lipomi intramidollari, per escludere l'eventuale presenza di siringomielia ed, in ultimo, per evidenziare le alterazioni microstrutturali del tronco cerebrale, non sempre visibili con le altre tecniche diagnostiche disponibili.

La valutazione del flusso mediante RM a contrasto di fase nel periodo pre-chirurgico è fondamentale, in quanto si dimostri essere un mezzo molto utile da utilizzare come guida nella pianificazione dell'intervento e per la previsione degli esiti chirurgici.

Per la CM I, attraverso la risonanza magnetica vengono ricercati i criteri diagnostici evidenziati nel documento di consenso, focalizzandosi in primis sul grado di erniazione delle tonsille cerebellari e dell'obex rispetto alla linea di McRae, il quale deve essere superiore a 5 mm.

Per la diagnosi e la caratterizzazione della Siringomielia, la RM con e senza l'utilizzo di mezzo di contrasto si conferma essere l'indagine radiologica più indicata, in quanto consenta una visualizzazione decisamente più accurata della siringa sul piano sagittale e assiale, rivelando facilmente la posizione, le dimensioni e l'estensione della cavità siringomielica.

La RM viene inoltre impiegata nel follow-up del paziente per documentare la storia naturale di tali patologie e per riconoscere tempestivamente una eventuale progressione nel corso di mesi o di anni.

Lo Studio del flusso di risonanza magnetica dinamica o CINE-MRI con gate cardiaco è un'indagine complementare utile all'analisi dell'idrodinamica del liquido cerebrospinale in modo non invasivo. Infatti, identifica facilmente le alterazioni della velocità e del flusso del liquido a livello del forame magno, visualizzandone il suo movimento nella parete del midollo e nella cavità siringomielica durante la sistole e la diastole cardiaca; un esame di questo tipo

risulta essere anche utile per documentare i cambiamenti del flusso post-operatorio ed i miglioramenti oggettivi.

La Mielografia con TC ad alta risoluzione è invece una scansione indicata in tutte quelle situazioni in cui non sia possibile utilizzare la risonanza magnetica come, ad esempio, nel caso di pazienti con impianti metallici nelle articolazioni o pazienti con pacemaker cardiaco.

Le scansioni TC ritardate possono visualizzare il colorante che si accumula nella cavità siringomielica ma, alcuni sostengono che però questa tipologia di esame abbia una bassa sensibilità nel rilevare il sito di ostruzione e che quindi sia poco utile nello studio della siringomielia associata alla Malformazione di Chiari di tipo I.

1.6 Trattamento chirurgico della Malformazione di Chiari I

Il trattamento chirurgico d'elezione Nella Malformazione di Chiari di tipo I è l'operazione di decompressione cranio-cervicale, che ha come scopo principale quello di ottenere le migliori condizioni spaziali nella transizione cranio-cervicale, incrementando quindi il volume della fossa posteriore e ricreando il passaggio interrotto o più semplicemente ridotto del liquido cerebro-spinale. Di conseguenza, utilizzando questa tecnica si alleggerisce la pressione esercitata sulle strutture nervose nel tentativo di interrompere la progressione della malattia ed anche di ridurre la gravità dei sintomi.

Questo tipo di intervento viene effettuato in anestesia totale con una durata circa pari ad un'ora e, nella maggior parte dei casi, ha un buon decorso operatorio; solitamente, infatti, non si ricorre ad alcuna terapia riabilitativa ed i sintomi dovuti alla fase successiva all'intervento che vengono più frequentemente segnalati sono il dolore nucale, la nausea e le vertigini nei primi giorni successivi all'operazione.

In alcuni soggetti però, è necessario ricorrere ad una decompressione con duraplastica espansiva in quanto, l'intervento di decompressione semplice precedentemente citato non risulti adeguato a garantire un corretto e sufficiente ampliamento della fossa cranica. In questi casi quindi, è sempre necessario ricorrere ad una plastica durale per il ripristino dell'anatomia. Questa procedura, come è facile dedurre, richiede più tempo rispetto alla semplice decompressione osteo- legamentosa ma, in compenso, presenta un outcome post-chirurgico decisamente migliore. In seguito a questo secondo tipo di intervento, il paziente viene mantenuto sotto osservazione per cinque giorni circa. Solitamente viene anche suggerito lo svolgimento di un follow up radiologico sfruttando la tecnologia di risonanza magnetica di

encefalo e midollo spinale per gli 8-10 mesi successivi all'operazione; inoltre, risulta anche molto utile l'analisi e la rivalutazione clinica periodica per accertare le condizioni del paziente, osservando se il soggetto abbia avuto un miglioramento, sia peggiorato oppure abbia mantenuto invariata la sua situazione clinica. Nel primo periodo successivo all'intervento è inoltre consigliata l'astensione dall'attività sportiva almeno fino allo svolgimento dei controlli successivi.

Come in ogni intervento chirurgico, ci sono delle possibili complicanze che è necessario tenere in conto. La complicanza operatoria più comune è la presenza di pseudomeningocele, ovvero un accumulo di liquido spinale sottomuscolare oppure sottocutaneo, che può, in alcuni casi, manifestarsi attraverso la fuoriuscita di liquido dalla ferita aperta ma che di solito è in grado di risolversi autonomamente. Inoltre, è possibile che si presentino altre eventuali complicanze, tra cui l'insorgenza di idrocefalo post-operatorio; lo sviluppo di idrocefalo in seguito all'operazione di decompressione di Chiari è una complicanza chirurgica nota, che riguarda una percentuale di soggetti compresa tra l'1% e il 9% dei pazienti operati.

L'idrocefalo e la Malformazione di Chiari di tipo I sintomatica però, presentano spesso manifestazioni simili della patologia, quindi, a causa della sovrapposizione dei sintomi e della variegata eziologia di queste due differenti patologie, è necessaria l'applicazione di un approccio clinico e chirurgico graduale e ponderato.

Il dolore post-operatorio è un ulteriore questione importante che riguarda i pazienti che hanno subito l'operazione di decompressione ed è solitamente gestito tramite l'utilizzo di oppioidi e miorilassanti; anche riguardo questo punto è necessario avere equilibrio ed attenzione perché, ad esempio, l'eccessiva sedazione attraverso gli oppioidi, può mettere il paziente a rischio di sviluppo di complicazioni ad essi legati, come la polmonite.

Per quanto riguarda il percorso neurochirurgico della siringomielia, è necessario sottolineare come spesso essa rappresenti una conseguenza di condizioni relative solamente al forame magno e per questa ragione, a seguito della decompressione della fossa cranica posteriore è possibile riscontrare un automatico miglioramento della sintomatologia legata alla formazione della cavità siringomielica, la quale può ridursi di dimensioni oppure si può stabilizzare spontaneamente: pertanto, un intervento apposito per la siringa di solito non è necessario.

1.7 Obiettivi

In seguito a tutte le informazioni fornite finora riguardanti la Malformazione di Chiari, nello specifico la CM I, è evidente che, per la patologia descritta non ci sia un percorso medico e chirurgico definito e sicuro.

Inevitabilmente un qualsiasi intervento, in particolare uno riferito al cervello ed eseguito in anestesia totale, porta con sé dei rischi intrinsechi, ma nel caso di questa anomalia, altre conseguenze derivano appunto dall'incertezza che tuttora è presente rispetto al percorso da seguire per ogni paziente.

Come è già stato precedentemente esposto, in generale l'intervento di decompressione osteo-legamentosa è il gold standard nella cura della Malformazione di Chiari, ma questa pratica non garantisce in tutti i casi l'ottenimento di una soluzione esaustiva al problema; spesso, infatti, questo tipo di intervento non solo non è in grado di fornire i necessari miglioramenti dello stato clinico del paziente, ma addirittura può causare ulteriori problemi, come per esempio la comparsa di siringomielia.

L'incertezza sull'esito di questo particolare intervento sarà reso chiaro dalla classificazione che verrà fornita in seguito: infatti, saranno inclusi nello studio solamente pazienti sottoposti alla decompressione della fossa cranica ed essi saranno suddivisi in primis in quattro classi differenti che denoteranno i possibili scenari che si devono affrontare; la suddivisione sarà basata sull'eventuale presenza di siringomielia precedente al trattamento chirurgico e sull'eventuale comparsa, soluzione o mantenimento della siringomielia in seguito all'intervento.

È immediatamente chiaro che sia necessario quindi andare a studiare i risultati ottenuti per provare a fornire degli strumenti robusti ed affidabili che possano guidare il medico verso il percorso di cura più adeguato per il singolo paziente: proprio questo è l'obiettivo della presente trattazione; utilizzando infatti differenti metodi di machine learning, ci si pone come fine quello di determinare, attraverso uno strumento di classificazione accurato, se il percorso seguito sia corretto e, in particolare, di capire se l'intervento di decompressione osteo-legamentosa sia davvero applicabile ad ogni soggetto in analisi.

Inoltre, la scelta della via di cura ad oggi è determinata dal valore di discesa tonsillare presente: considerando che spesso i risultati non siano quelli sperati, ci si soffermerà anche sulla base su cui le decisioni si fondano, andando quindi ad analizzare quali possano essere le informazioni più utili per selezionare la giusta strada da percorrere concentrandosi nello specifico sui parametri morfologici estratti dalle immagini di RM e sui sintomi presenti alla diagnosi.

2. Metodo

2.1 Descrizione generale

L'obiettivo principale di questo lavoro è quello di valutare l'influenza di differenti fattori nella diagnosi e nella scelta di cura della malformazione di Chiari.

In particolare, si è scelto di costruire diversi classificatori e diverse reti neurali con caratteristiche eterogenee e modificare solamente i dati di input:

- nel primo caso si considerano solamente i parametri morfologici estratti dalle immagini;
- nel secondo caso si considerano solamente i sintomi preoperatori estratti dalle cartelle cliniche;
- nel terzo caso l'input sarà composto dai precedenti dati uniti insieme.

Confrontando queste tre casistiche applicate alle differenti strutture che verranno considerate ed analizzandone i risultati ottenuti si andrà a comprendere quali informazioni (o quale combinazione di essi) possano essere più adatte al miglioramento della valutazione generale del paziente.

2.2 Estrazione dei parametri morfologici

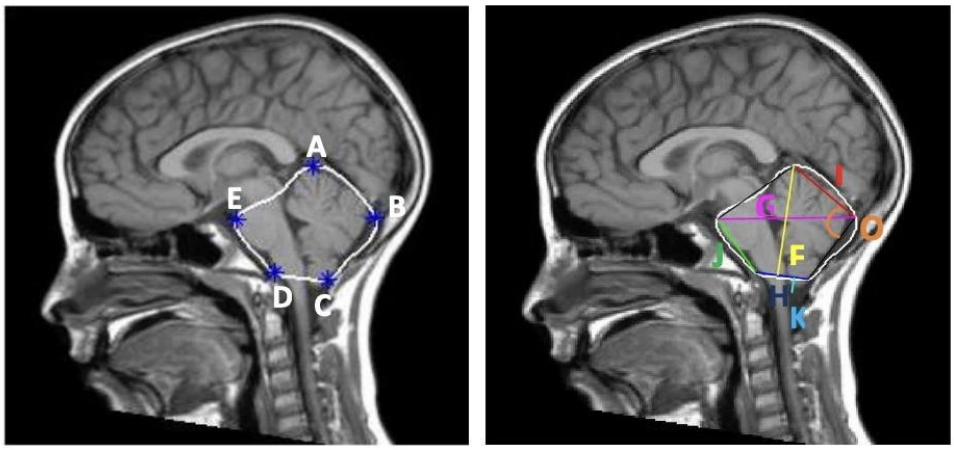
Dopo aver ricavato automaticamente le segmentazioni di cervello, fossa posteriore e cervelletto tramite il metodo dei demons, è possibile procedere all'estrazione dei parametri morfologici d'interesse, attraverso la costruzione di un pentagono dal quale vengono ricavate le quindici misurazioni considerate.

Come è stato descritto nel paragrafo di diagnosi della CM-I, un parametro di fondamentale importanza è la lunghezza dell'erniazione tonsillare. Infatti, il criterio più comunemente utilizzato nella diagnosi della CM-I è attualmente basata sull'osservazione radiografica della lunghezza dell'erniazione tonsillare, la quale dev'essere superiore a 5 mm perché venga determinata la presenza della malformazione di Chiari. Questo perché si assume che l'ernia tonsillare sia una conseguenza della costrizione craniale dovuta alla fossa posteriore ipoplastica. Però molti sintomi che si sono osservati non hanno alcuna correlazione con la presenza dell'ernia e quindi la lunghezza dell'ernia tonsillare non è necessariamente correlata con la gravità dei sintomi o la scelta del trattamento chirurgico. (Urbizu, 2017)

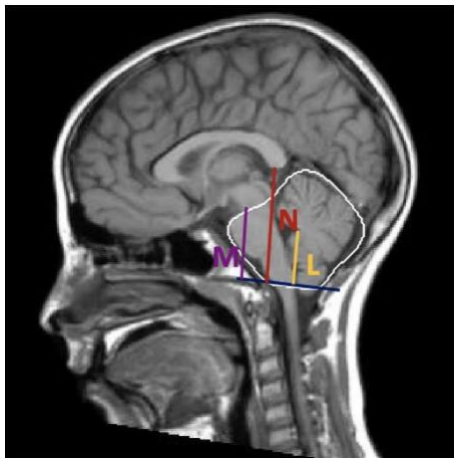
Pertanto, negli ultimi anni, sono stati condotti differenti studi per identificare ulteriori parametri che possano migliorare la diagnosi.

In questa trattazione sono stati considerati alcuni dei parametri descritti in (Urbizu, 2017) e (Urbizu A, 2014), di cui la maggior parte sono stati calcolati a partire dal valore relativo all'area della fossa posteriore. Per questa ragione, infatti, è stato sviluppato un algoritmo in grado di calcolare i vertici di un pentagono che approssimi la sagoma della fossa posteriore.

Dal pentagono ricavato vengono calcolati la posizione del tentorio (A), del basion (D) e dell'opsthion (C), l'altezza della fossa posteriore (F), il diametro antero-posteriore della fossa posteriore (G), la protuberanza occipitale (B), la lunghezza del tentorio (I), la lunghezza del clivus (J), la posizione del piano del forame magno (H) e l'angolo del tentorio (O).



Considerato il piano del forame magno (H) viene calcolata la lunghezza ernia tonsillare (K). Sono state considerati anche altri parametri geometrici come distanza fastigio-forame magno (L), distanza pons-forame magno (M), distanza corpo calloso-forame magno (N) che però sono stati calcolati manualmente da un operatore medico.



2.3 Valutazione dei sintomi

Per l'ottenimento dei sintomi è stata fondamentale la collaborazione con la Facoltà di medicina e chirurgia dell'Università di Torino, dalla quale è stata svolta un'analisi dal punto di vista clinico. È necessario quindi introdurre alcune informazioni fondamentali fornite dalla collega che ha condotto tale studio.

In particolare, sono state combinate le esigenze ingegneristiche come, ad esempio, la qualità dell'immagine radiologica fornita ed i parametri medici scelti, di seguito elencati.

Sono stati arruolati nello studio tutti i pazienti che soddisfino i seguenti criteri:

- Pazienti affetti da Malformazione di Arnold-Chiari I con o senza siringomielia e sottoposti ad intervento chirurgico di decompressione della fossa cranica posteriore;
- Erniazione di almeno 1 tonsilla cerebellare ≥5 mm al di sotto della linea di McRae (basion-to-opisthion) su immagini sagittali pesate in T1 e almeno uno dei seguenti reperti clinici: cefalea posteriore peggiorata con la manovra di Valsalva, disturbi dei nervi cranici misti (disfagia, disfonia, singhiozzo), disturbi cranici oto-vestibolari (vertigini, acufeni) disturbi delle vie lunghe (problemi motori e sensitivi), segni cerebellari (atassia, dismetria, tremori), disturbi sfinterici e scogliosi;
- Pazienti sottoposti ad intervento di decompressione osteo-legamentosa oppure di decompressione della fossa posteriore con plastica durale con o senza coartazione tonsillare;
- periodo dell'intervento chirurgico compreso tra gennaio 2010 e dicembre 2020;
- presenza di dati clinici e strumentali di follow-up a 1 anno;
- firma del consenso informato all'intervento e all'uso dei dati clinici a scopo di ricerca.

Non sono stati arruolati nello studio i pazienti che abbiano sviluppato siringomielia prima dell'intervento non correlata ad anomalia di Chiari e quelli che sono stati sottoposti ad altri tipi di intervento chirurgico.

2.4 Struttura della classificazione

Di seguito vengono riportate le caratteristiche comune a tutte le classificazioni selezionate, che, come obiettivo, hanno la valutazione dell'influenza dei parametri morfologici e dei sintomi del paziente nella determinazione del percorso medico e chirurgico da seguire.

2.4.1 Organizzazione dei dati

Il dataset è composto da 55 pazienti: questi sono stati selezionati attraverso le esigenze ingegneristiche e mediche combinate insieme, quindi seguendo i parametri di inclusione precedentemente elencati.

Il dataset è stato poi suddiviso in:

- Training set, utile per allenare la rete e composto da 44 pazienti, quindi l'80% del dataset completo e selezionati utilizzando una estrazione random proporzionale alla classe.
- Validation set, utile per validare la rete e composto da 11 pazienti, quindi il 20% del dataset completo e selezionati utilizzando una estrazione random proporzionale alla classe.

Vista però la ridotta quantità di pazienti inclusi nello studio e di conseguenti immagini disponibili, la fase di validazione è stata svolta utilizzando il metodo chiamato "Leave One Out"; in tale procedura, ciclicamente, tutti i campioni tranne uno vengono utilizzati per comporre il dataset utilizzato per addestrare la rete ed il paziente escluso verrà utilizzato per validarla. In questo modo tutti i pazienti vengono utilizzati sia per addestrare la rete che per validarla, in modo che la costruzione del modello non sia eccessivamente condizionata dai dati utilizzati per le due diverse fasi.

Per valutare le differenti influenze dei parametri morfologici e dei sintomi nella valutazione del trial medico più adeguato, si è scelto di svolgere tre classificazioni differenti nelle quali cambiassero però soltanto le informazioni utilizzate come input e mantenendo l'output sempre uguale: nel primo caso si considerano solamente i parametri morfologici estratti manualmente ed automaticamente dall'immagine di risonanza magnetica; nel secondo caso si considerano solamente i sintomi estratti dalle cartelle cliniche dei pazienti mentre, nel terzo caso, si uniscono i due differenti input precedenti per osservare la loro combinazione.

L'obiettivo primario è lo studio delle differenze per poter trarre delle conclusioni in merito alla rilevanza che questi parametri possano avere nella previsione del decorso del paziente.

I parametri morfologici estratti automaticamente sono:

- area del cervello;
- area della fossa posteriore;

- area del cervelletto;
- rapporto tra area del cervelletto e della fossa posteriore;
- rapporto tra area della fossa posteriore e del cervello lunghezza del tentorio;
- lunghezza del clivus;
- diametro antero-posteriore;
- altezza fossa posteriore;
- ernia tonsillare;
- angolo del tentorio;
- lunghezza forame magno.

Inoltre, vengono aggiunti parametri morfologici estratti manualmente dai medici, quali:

- distanza pons-forame magno;
- distanza corpo calloso-forame magno;
- distanza fastigio-forame magno.

Gli altri dati di input saranno invece composti dai seguenti sintomi estratti dalle cartelle cliniche:

- cefalea nucale;
- disturbi dei nervi cranici misti (disfagia, disfonia, singhiozzo);
- disturbi dei nervi cranici oto-vestibolari (vertigini, acufeni);
- disturbi delle vie lunghe (problemi motori/sensitivi);
- segni cerebellari (atassia, dismetria);
- disturbi sfinterici;
- scoliosi;
- altro.

2.4.2 Classi

Per giustificare la scelta fatta per la determinazione delle classi, è necessario introdurre un'altra porzione fondamentale del lavoro svolto dall'Università degli studi di Torino.

In particolare, l'obiettivo principale di questa trattazione è quello di stimare l'associazione tra specifici parametri morfologici del cranio estratti dalle immagini di risonanza magnetica sagittale T1 pesata di pazienti operati per anomalia di Chiari I ed il rischio di un peggioramento clinico-radiologico entro il primo anno dall'intervento; inoltre, l'obiettivo secondario consiste nello stabilire quali misure morfometriche predicano con più accuratezza il rischio di sviluppo

della siringomielia post- chirurgica, nonché di selezionare il tipo di intervento più appropriato, sulla base dei criteri radiologici estrapolati dalla RM encefalica. Per fare ciò, quindi, la dottoressa ha deciso di sviluppare un'analisi statistica, dividendo i pazienti in 4 differenti gruppi:

- gruppo 0: comprende pazienti sintomatici, senza siringomielia alla diagnosi, i quali a seguito dell'intervento chirurgico, hanno avuto un miglioramento clinico e radiologico entro il primo anno, ovvero hanno assistito ad una riduzione della sintomatologia e dell'ernia tonsillare;
- gruppo 1: include pazienti sintomatici, con siringomielia alla diagnosi che, dopo l'intervento chirurgico, hanno avuto un miglioramento clinico e radiologico al primo anno, con una riduzione della cavità siringomielica;
- gruppo 2: raggruppa pazienti sintomatici con siringomielia in partenza i quali, successivamente all'intervento chirurgico, hanno avuto un peggioramento clinico e radiologico, con un incremento delle dimensioni della cavità siringomielica oppure con una persistenza dei sintomi legati alla presenza di siringomielia;
- gruppo 3: identifica pazienti sintomatici nonché privi di cavità siringomielica alla diagnosi che, in seguito all'intervento di decompressione, hanno sviluppato la siringomielia entro il primo anno.

Possiamo notare come questa classificazione sia caratterizzata da una componente fondamentale, cioè la presenza di siringomielia antecedente o successiva all'intervento; questo parametro è infatti ciò su cui è costruita l'analisi statistica precedentemente citata.

Inoltre, è necessario specificare che il tipo di intervento citato nella suddivisione del dataset sopra riportata è la decompressione osteo-legamentosa o di plastica durale con o senza coartazione tonsillare perché, secondo le attuali linee guida, queste procedure sono considerate come trattamenti di prima scelta nel soggetto pediatrico affetto da Malformazione di Chiari sintomatica.

Queste informazioni sono state utilizzate per decidere il tipo di classificazione che si volesse ottenere: infatti, si è deciso insieme di mantenere una coerenza tra i due differenti elaborati, avendo essi obiettivi molto affini.

Partendo quindi dalla suddivisione sopra mostrata si è deciso di utilizzare classificatori binari; nello specifico la classificazione effettuata prevede tale suddivisione:

- classe 1: è composta dall'unione del gruppo 0 e del gruppo 1 precedentemente delineati. Questa classe, quindi, rappresenterà i pazienti che abbiano effettivamente tratto giovamento dall'intervento di decompressione osteo-legamentosa;
- classe 0: è composta dall'unione del gruppo 2 e del gruppo 3 precedentemente delineati. Questa classe, quindi, rappresenterà i pazienti che non abbiano effettivamente tratto giovamento dall'intervento di decompressione osteo-legamentosa o che addirittura siano stati da esso penalizzati.

Da questa classificazione si può dedurre come sarà svolta l'analisi dei risultati ottenuti.

Fino ad ora ci si è basati su una soglia numerica applicata ad un parametro morfologico per decidere se attuare o meno la decompressione osteo-legamentosa; il nuovo obiettivo sarà quindi quello di rispondere ad una domanda: è possibile, tramite il machine learning, determinare quale tipo di intervento sia più adeguato ad un determinato paziente?

Inoltre, si cercherà di comprendere e determinare quali siano i parametri più rilevanti ed influenti nella scelta della corretta procedura.

2.4.3 Classificatori

Come già detto in precedenza, l'analisi di questa trattazione è totalmente basata sul confronto di tre differenti casi la cui unica differenza è individuabile nel dataset utilizzato come input:

I. Il primo caso vedrà come dati di input solamente 15 parametri morfologici, cioè misure geometriche estratte direttamente dalle immagini di risonanza magnetica; l'estrazione avviene in parte manualmente, come ad ora è consuetudine fare, ed in parte attraverso il metodo automatico precedentemente trattato.
II. Il secondo caso vedrà come dati di input solamente 8 sintomi estratti direttamente dalle cartelle cliniche dei pazienti; ovviamente è necessario mantenere l'anonimato dei partecipanti e, per questa ragione, è stato creato un codice identificativo che permettesse di mantenere la privacy dei pazienti e che è stato poi esteso anche per la distinzione delle immagini di risonanza degli stessi.
III. Il terzo caso vedrà come dati di input la combinazione dei due precedenti dataset e servirà principalmente da riferimento: infatti, sarebbe stato limitante confrontare semplicemente le due reti precedenti tra loro; l'idea, quindi, è di introdurre l'unione di

esse per verificare l'effettiva influenza dei differenti parametri sulla riuscita di una corretta classificazione.

Il primo strumento utilizzato per osservare la qualità della classificazione è una rete neurale: per non favorire nessuna tipologia di dataset e per garantire di avere coerenza tra i risultati, ovviamente si è scelto di mantenere le stesse identiche condizioni per tutte e tre le casistiche.

Innanzitutto, la classificazione sarà ovviamente supervisionata: nella fase di allenamento quindi, oltre alle informazioni relative ai differenti parametri scelti come dati di input, sarà inserito anche un ulteriore vettore di input che conterrà le classi reali di appartenenza dei pazienti; questa struttura è necessaria per poter attuare una classificazione ed è quindi fondamentale, in fase di allenamento, fornire alla rete un riferimento per quanto riguarda la classe di appartenenza dei campioni.

La seconda scelta da fare è relativa proprio al tipo di rete da utilizzare: in questo caso viene usata una rete di tipo feed-forward; questa tipologia è la più semplice a livello di funzionamento, ma è comunque in grado di fornire risultati utili ad un'analisi generale del problema. Le reti feed-forward, come è reso chiaro proprio dal nome, si distinguono dalle altre per il fatto che le informazioni si muovano solamente "in avanti": le connessioni tra i nodi non formano cicli, ma le informazioni si muovono soltanto in un'unica direzione rispetto ai nodi d'ingresso ed attraverso nodi nascosti (se presenti) fino ai nodi d'uscita.

È necessario poi selezionare la modalità di analisi dei risultati ottenuti in seguito all'allenamento della rete ed alla sua successiva validazione: in questo caso il mezzo scelto per osservare il funzionamento e l'affidabilità delle differenti reti è la confusion matrix (la cui struttura verrà dettagliatamente chiarita in seguito), un mezzo molto classico ma molto utile e chiaro per valutare la qualità della classificazione ottenuta.

La scelta dell'operatore, quindi, è semplicemente riguardante l'eventuale presenza e di conseguenza il numero di strati nascosti che si vogliono inserire ed il numero di neuroni che compongono ogni strato.

In questo caso si è scelto di utilizzare un numero moderatamente alto di neuroni per ogni strato nascosto, nella speranza che, pur avendo relativamente pochi dati (solo 55 pazienti hanno rispettato tutti i criteri di inclusione), la creazione di molte connessioni potesse migliorare l'output; vengono quindi elencate le 4 strutture che verranno confrontate tra loro:

- Rete 1: nella prima rete vengono inseriti 3 layers nascosti, rispettivamente composti da 30, 20 e 10 neuroni ciascuno.

- Rete 2: nella seconda rete vengono inseriti 3 layers nascosti, rispettivamente composti da 20, 10 e 5 neuroni ciascuno.
- Rete 3: nella terza rete viene sostanzialmente "raddoppiata" la rete 1, quindi vengono inseriti sei layers nascosti; i primi due saranno composti da 30 neuroni, il terzo ed il quarto strato saranno composti da 20 neuroni ciascuno e gli ultimi due strati saranno composti da 10 neuroni.
- Rete 4: nella terza rete viene sostanzialmente "raddoppiata" la rete 2, quindi vengono inseriti sei layers nascosti; i primi due saranno composti da 20 neuroni, il terzo ed il quarto strato saranno composti da 10 neuroni ciascuno e gli ultimi due strati saranno composti da 5 neuroni.

Queste 4 reti serviranno per il primo confronto generico: per ognuna delle reti elencate verrà svolto un allenamento ovviamente solo sul dataset di training; tale allenamento subirà 100 ripetizioni per ognuna delle reti costruite e da questa reiterazione verranno estratte delle confusion matrix contenenti i valori medi delle 100 prove. Tale procedimento verrà svolto ovviamente per ognuna delle 3 casistiche considerate (solo parametri morfologici, solo sintomi ed entrambi unitamente).

Al termine di questo processo quindi, si otterranno 4 confusion matrix (contenenti i valori medi delle 100 ripetizioni) per ognuno dei tre differenti scenari: queste quattro matrici verranno confrontate tra loro e, sulla base di parametri estraibili (come ad esempio l'accuratezza, la sensibilità e la specificità) verrà selezionata la rete più adeguata allo scopo.

Solo successivamente, in seguito alla scelta della rete, verrà nuovamente svolta la fase di allenamento ed in questo caso anche la fase di validazione della rete.

Infine, per ognuna delle tre casistiche considerate, sarà ottenuta una matrice relativa alla fase di allenamento ed una matrice relativa alla fase di validazione; queste sei matrici verranno quindi confrontate tra loro per poter selezionare la modalità di classificazione migliore.

Il secondo metodo di classificazione è attuato attraverso l'utilizzo di un classificatore di tipo SVM (Support Vector Machine): in generale, i modelli di questo tipo hanno come obiettivo quello di trovare la retta di separazione tra le classi che massimizza il margine tra di esse; il margine rappresenta quindi la distanza minima dalla retta degli elementi appartenenti alle due classi. Questo obiettivo viene raggiunto utilizzando una piccola parte del dataset di addestramento: infatti solo i cosiddetti vettori di supporto vengono presi d'esempio per la

costruzione del modello. Questi elementi sono proprio quelli che giacciono sul margine, e sono quindi i pazienti appartenenti ad una classe ma che più si avvicinano all'altra; sono quindi gli elementi più difficili da classificare. Utilizzare questi campioni significa iniziare la classificazione proprio dai dati che definiranno la retta di separazione ed il margine, rendendo così ininfluenti tutti gli altri elementi più facili da classificare.

Il terzo metodo di classificazione utilizzato è il k-NN (K Nearest Neighbors): in generale, fornito un nuovo dato d'ingresso da classificare, il modello cerca un certo numero di campioni all'interno del dataset più vicino all'input in questione; la variabile k è proprio il parametro che determina quanti esempio simili sia necessario trovare. Infine, la classe di appartenenza assegnata al dato di ingresso sarà quella è maggiormente rappresentata dagli esempi trovati. Inoltre, il secondo parametro che influisce sul risultato ottenuto è la metrica selezionata per calcolare la distanza tra gli elementi.

Il quarto classificatore è quello denominato Naive-Bayes, il quale crea un modello di classificazione probabilistica della relazione tra un insieme di variabili predittive ed una variabile target; il classificatore Naive-Bayes calcola quindi la probabilità della variabile target di appartenere a ciascuna delle due classi considerate. Questo modello è definito come un classificatore bayesiano "ingenuo" (traduzione appunto di "Naive") oppure semplificato viste le semplici ipotesi di partenza; è infatti basato su un modello di probabilità che ipotizza che ci sia indipendenza tra le caratteristiche, cioè assume che la presenza o l'assenza di una particolare caratteristica non sia correlata alla presenza o all'assenza di altre informazioni.
Per calcolare la probabilità di appartenenza ad una classe, questo strumento usa ovviamente il teorema di Bayes, che afferma che la probabilità che un dato elemento appartenga alla classe X considerando le informazioni riguardo la classe Y è uguale al rapporto tra:

- il prodotto tra la probabilità che il dato elemento appartenga alla classe Y considerando le informazioni riguardo la classe X e la probabilità a priori che il dato elemento appartenga alla classe X senza considerare le informazioni riguardo la classe Y;
- la probabilità che il dato elemento appartenga alla classe Y senza considerare le informazioni riguardo la classe X.

2.4.4 Applicazione della feature selection

In seguito all'elaborazione dei dati precedentemente descritti, si è quindi deciso di applicare la feature selection: quando risulta necessario riconoscere un determinato pattern oppure elaborare immagini (in questo caso di natura medica e diagnostica) la selezione delle caratteristiche è una metodologia atta a ridurre la dimensionalità del dataset in esame.

La selezione delle caratteristiche, quindi la riduzione dei dati di ingresso, è necessaria per individuare le informazioni maggiormente significative al fine di una classificazione.

Questo processo risulta inoltre adeguato alla creazione di un modello funzionale in cui si vogliano ridurre le dimensioni dei dati di input, imponendo quindi una soglia minima al numero di informazioni da considerare durante la creazione di una data rete.

In aggiunta, spesso i dataset utilizzati per la costruzione di un modello predittivo presentano caratteristiche ridondanti, quindi ripetitive o comunque futili al miglioramento della classificazione; la feature selection poi, è fondamentale nel caso in cui ci sia bisogno di rendere più efficiente il processo in corso, ad esempio diminuendo il suo peso computazionale a livello di CPU e di memoria.

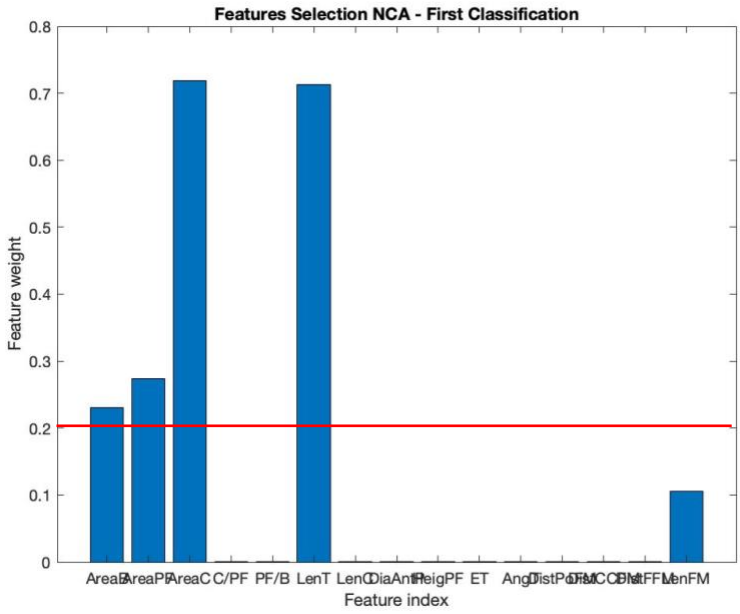
Infine, la selezione delle caratteristiche può anche essere utile per marginare il problema dell'overfitting, quindi per evitare che una rete sia troppo addestrata a riconoscere i dati di input utilizzati per l'addestramento ma che poi questa non sia in grado di funzionare altrettanto adeguatamente rispetto alla validazione ed eventualmente alla fase di test.

Ovviamente ci sono differenti modalità di attuazione della feature selection, ma in questa trattazione ne sono state considerate nello specifico le seguenti due:

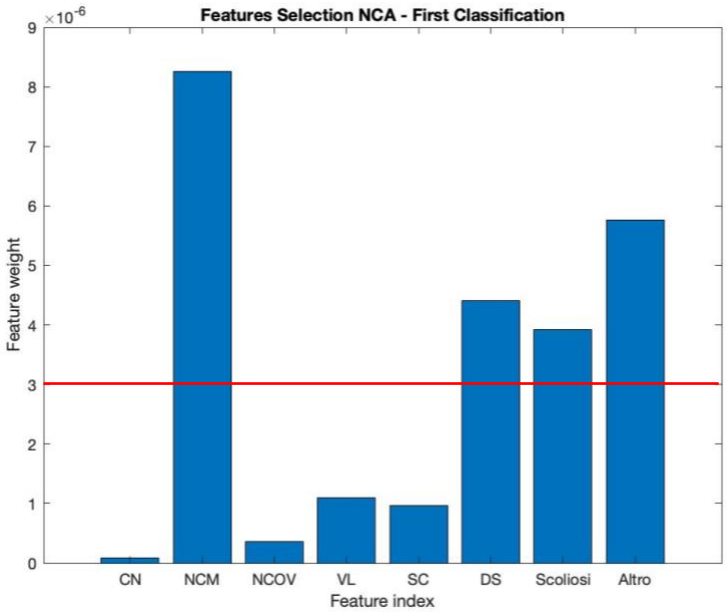
- Neighborhood component analysis (NCA): è un metodo non parametrico per selezionare le caratteristiche nel caso in cui si voglia massimizzare l'accuratezza di una determinata predizione; in questo caso vengono analizzati i pesi delle caratteristiche con lo scopo di minimizzare una funzione oggetto che misura la classificazione media su tutto il dataset di addestramento.
- Massima rilevanza e minima ridondanza (MRMR): è un metodo che, oltre ad analizzare la massima rilevanza delle features di ingresso, processo che viene sempre attuato nel corso della selezione delle caratteristiche attraverso, ad esempio, alla minimizzazione della varianza, è anche in grado di ricercare il livello minimo di ridondanza delle informazioni.

Di seguito si riportano le immagini ottenute successivamente all'applicazione della feature selection; le prime mostreranno i risultati ricavati in seguito all'utilizzo della modalità NCA applicata sia ai 15 parametri morfologici sia agli 8 sintomi dei pazienti.

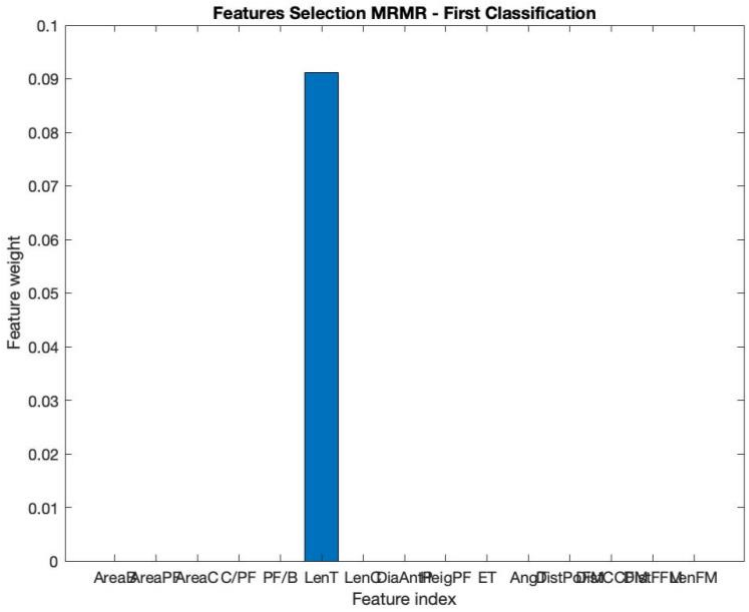
Applicando la prima tipologia di feature selection ai parametri morfologici è possibile notare come le caratteristiche considerate come più rilevanti siano quelle relative all'area del cervello, del cervelletto, della fossa cranica ed alla lunghezza del tentorio.



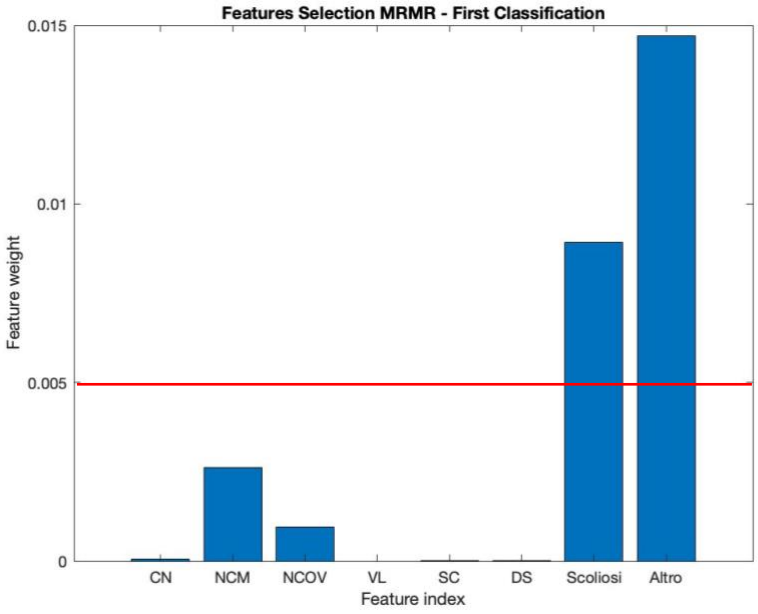
Applicando la prima tipologia di feature selection ai sintomi diagnostici è possibile notare come le caratteristiche considerate come più rilevanti siano quelle relative alla presenza preoperatoria di disturbi dei nervi cranici misti, di disturbi sfinterici di scoliosi e di altri differenti sintomi che possono presentarsi in fase diagnostica.



Applicando il secondo metodo per selezionare le caratteristiche ai parametri morfologici, quindi la MRMR, non si ottengono informazioni affidabili, in quanto, l'unica feature considerata più rilevante e minimamente ridondante è quella relativa alla lunghezza del tentorio.



In seguito all'applicazione delle MRMR ai sintomi diagnostici emerge che i dati più rilevanti siano relativi solamente alla presenza preoperatoria di scoliosi e di altri eventuali sintomi che possano essere riscontrati in fase di diagnosi.



In seguito alle immagini prodotte dai due differenti metodi di feature selection, la scelta del metodo da utilizzare ricade facilmente sul metodo della Neighborhood component analysis (NCA): come già precedentemente detto, la MRMR applicata ai parametri morfologici estratti dalle immagini di risonanza magnetica restituisce un'unica caratteristica come rilevante e questo porterebbe sicuramente all'ottenimento di risultati distorti; inoltre, in seguito all'applicazione della MRMR ai sintomi presenti alla diagnosi, l'output include tutti i valori dapprima considerati, quindi non si corre alcun rischio di perdere informazione e qualità.

In conclusione, i dati che comporranno l'input successivamente all'utilizzo della feature selection secondo saranno:

- Area del cervello;

- Area del cervelletto;
- Area della fossa cranica posteriore;
- Lunghezza del tentorio.

Inoltre, i sintomi diagnostici che verranno utilizzati per comporre il dataset saranno:

- Disturbi dei nervi cranici misti;
- Disturbi sfinterici;
- Scoliosi;
- Altri sintomi presenti alla diagnosi.

Quindi, l'applicazione della feature selection significherà semplicemente considerare soltanto le informazioni appena elencate: ovviamente, nel caso della prima rete verranno incluse solamente le quattro caratteristiche relative ai parametri morfologici, nella costruzione della seconda rete verranno incluse le quattro caratteristiche relative ai sintomi e l'unione di essi verrà utilizzata per lo sviluppo e la validazione della terza ed ultima rete.

3. Risultati delle reti neurali

In questo capitolo verranno riportati i risultati ottenuti in seguito all'applicazione delle reti precedentemente descritte. È necessario chiarire i metodi di analisi che verranno utilizzati: innanzitutto ogni rete sarà contraddistinta da una relativa confusion matrix.

La confusion matrix è uno strumento molto utile per analizzare la qualità di una rete neurale sotto vari aspetti, come ad esempio in ambito di sensitività e di specificità, fattori che di seguito verranno considerati e descritti più nel dettaglio.

	CLASSE REALE		
CLASSE PREDETTA		Positivo	Negativo
	Positivo	VP	FP
	Negativo	FN	VN

La prima precisazione necessaria è relativa alla distinzione tra le classi: come da convenzione la classe 0 è la classe negativa, in questo caso tutti i pazienti che hanno riscontrato problematiche in seguito all'intervento mentre, la classe 1 è la classe positiva, quindi quella che include i pazienti che non hanno riscontrato peggioramenti di alcun tipo in seguito all'operazione.

Nella confusion matrix le combinazioni prendono forma in base alla classe reale di appartenenza degli elementi considerati ed alla classe di appartenenza assegnata dalla rete in questione: sostanzialmente vengono confrontate le classi di input e le classi di output.

Inoltre, è possibile notare quattro differenti quadranti:

I. Veri positivi (VP): sono tutti i pazienti appartenenti alla classe 1 che vengono poi considerati dalla rete come appartenenti effettivamente alla classe 1; in questo caso, quindi sono tutti quei pazienti per cui l'intervento è risultato adeguato che vengono riconosciuti come tali.
II. Falsi positivi (FP): sono tutti i pazienti realmente appartenenti alla classe 0 negativa, quindi, coloro per i quali l'intervento non risulterebbe giovante, ma che erroneamente vengono classificati in classe 1 positiva; quindi, considerati dalla rete come pazienti per i quali l'intervento sarebbe adatto, nonostante ciò non rispecchi la realtà.
III. Falsi negativi (FN): sono tutti i pazienti realmente appartenenti alla classe 1 positiva, quindi, coloro per i quali l'intervento risulterebbe giovante, ma che erroneamente vengono classificati in classe 0 negativa; quindi, considerati dalla rete come pazienti per i quali l'intervento non sarebbe adatto, nonostante ciò non rispecchi la realtà.
IV. Veri negativi (VN): sono tutti i pazienti appartenenti alla classe 0 che vengono poi considerati dalla rete come appartenenti effettivamente alla classe 0; in questo caso, quindi sono tutti quei pazienti per cui l'intervento non è risultato adeguato che vengono riconosciuti come tali.

Nel momento in cui si sceglie di utilizzare la confusion matrix come mezzo per analizzare la qualità della rete, non solo si vogliono ottenere i risultati migliori rispetto alla classificazione corretta, ma è necessario rispondere ad un quesito fondamentale: è peggio avere un numero maggiore di falsi positivi oppure di falsi negativi?

In questo caso, quindi, si dovrà decidere se considerare più sconveniente il caso in cui venga suggerito ed eventualmente eseguito un intervento che non gioverà al paziente o che, peggio ancora, possa provocargli dei danni, oppure se considerare peggiore il caso in cui non venga attuata un'operazione che potrebbe essere di grande aiuto.

Considerato il fatto che la Malformazione di Chiari abbia un tasso di mortalità molto basso e considerando anche che un'operazione al cervello porti con sé i rischi legati ad un intervento ed all'anestesia, in questa trattazione viene scelto di valutare migliori i casi in cui il numero di falsi positivi sia il minore possibile.

Da questa tabella poi verranno estratti due valori fondamentali in un test diagnostico: in un test diagnostico, la sensibilità è il rapporto tra gli elementi classificati come veri positivi e tutti i campioni realmente positivi ed è una misura di quanto bene un test possa identificare i veri positivi e la specificità è il rapporto tra gli elementi classificati come veri negativi e tutti i campioni realmente negativi ed è una misura di quanto bene un test possa invece identificare i veri negativi. Per tutti i test, sia diagnostici che di screening, di solito c'è un compromesso tra sensibilità e specificità, tale che sensibilità più elevate significheranno specificità inferiori e viceversa.

Inoltre, è possibile valutare l'accuratezza della rete, una misura più generale che indichi l'affidabilità della classificazione: in particolare indica quanto la rete sia in grado di classificare correttamente gli elementi di input ed è definita come il rapporto tra la somma degli elementi correttamente classificati ed il numero totale di campioni.

3.1 Classificazione con i parametri morfologici

I primi risultati riportati sono quelli relativi alla rete che, come input ha solamente i 15 parametri morfologici estratti dalle immagini radiografiche.

Come anticipato in precedenza, il primo step è riferito ad una media fatta su 100 ripetizioni: viene ripetuta la fase di allenamento della rete; quindi, l'input sarà composto solamente dal training set, ed ogni valore di output viene mediato con i precedenti.

RETE 1	
9.88	6.64
6.12	21.36

RETE 2	
9.51	4.86
6.49	23.14

RETE 3	
8.52	4.65
7.48	23.35

RETE 4	
8.12	4.75
7.88	23.25

Come spiegato in precedenza, questi valori sono il risultato della media di 100 ripetizioni applicate ad ogni rete: questo processo risulta molto utile per confrontare in modo netto le quattro reti scelte e di conseguenza per selezionare il metodo di classificazione più adatto allo scopo.

È immediato osservare che le reti diano risultati abbastanza simili tra loro, ma si possono comunque osservare alcuni fenomeni. Precedentemente, nell'introduzione al seguente capitolo, sono stati esplicati alcuni criteri utili per determinare la rete a cui riferirsi: ovviamente bisogna preferire la rete con il maggior numero di campioni, in questo caso pazienti,

correttamente classificati però è necessario anche favorire la rete con il minor numero possibile di falsi positivi. In questo caso l'accuratezza è visibilmente migliore in seguito all'utilizzo della seconda rete e sarà proprio questa struttura ad essere utilizzata: infatti, nonostante il numero di falsi positivi non sia il minore, comunque si discosta di poco dal valore minimo ottenuto; inoltre, selezionando la seconda rete si ha anche un buon equilibrio tra i due differenti tipi di pazienti erroneamente classificati, senza così avere un valore riferito ai campioni considerati falsi negativi eccessivamente alto, nonostante questa non sia la condizione peggiore.
Dopo aver scelto la rete quindi, si applicherà la medesima struttura: la prima fase sarà quella di allenamento, svolta sul training set; in questo caso però sarà svolta la fase di validazione per osservare come la rete sia in grado di classificare elementi sconosciuti.

Training set	
12	6
8	28

Validation set	
9	10
11	25

Per quanto riguarda la fase di addestramento della rete si possono notare risultati accettabili: nessuno dei tre valori calcolati risulterà eccezionalmente elevato, però comunque presenteranno tutti valori con cui poter eventualmente lavorare. Nella fase di validazione però non possono valere le stesse considerazioni; infatti, i pazienti che vengono classificati correttamente diminuiscono e, fattore ancor più grave, il numero di pazienti realmente positivi e classificati come positivi è minore del numero di pazienti realmente positivi ma classificati come appartenenti alla classe negativa; da ciò è facile dedurre che il valore di sensibilità sarà inferiore al 50% e ciò indicherà che questo specifico test non sia per nulla affidabile per quanto riguarda la classificazione dei pazienti per i quali l'intervento sarebbe effettivamente giovante. E' necessario specificare che una situazione di questo tipo non sia vantaggiosa: infatti, in fase di costruzione di una qualsiasi rete, si vuole evitare di ottenere buoni risultati relativi all'addestramento per poi notare un peggioramento in fase di validazione della rete; questo risultato indica che la rete sia stata eccessivamente adattata ai dati di training senza poi essere in grado di riconoscere nuovi dati su cui non è stata allenata, che in fondo è proprio il compito principale di un algoritmo di classificazione: questo fenomeno prende il nome di overfitting.

3.1.1 Applicazione della feature selection

RETE 1	
8.89	4.99
7.11	23.01

RETE 2	
8.05	5.36
7.95	22.64

RETE 3	
7.53	3.68
8.47	24.32

RETE 4	
6.59	4.56
9.41	23.44

In seguito all'applicazione della feature selection, per quanto riguarda la fase di confronto delle 4 differenti strutture, i risultati ottenuti non variano in maniera: l'accuratezza, quindi il numero di corretti classificati, rimane circa simile al caso precedente mentre invece, il numero di campioni erroneamente classificati come positivi rimane circa costante, restituendo però un valore minimo inferiore rispetto alla classificazione precedente, come si può notare nella terza rete.

In questo caso, seguendo nuovamente i parametri elencati in precedenza, la scelta della rete avviene come segue: i valori di accuratezza sono molto simili nella prima rete e nella terza, ma il valore relativo ai falsi positivi è nettamente inferiore in quest'ultima. Nonostante, scegliendo la rete più semplice si avrebbe un maggior equilibrio tra i pazienti erroneamente classificati, la differenza riguardo i falsi positivi, situazione che è stata considerata essere la peggiore, è troppo elevata per non essere considerata; la rete utilizzata sarà quindi la terza.

Training set	
8	0
12	34

Validation set	
8	8
12	27

In questo caso si possono osservare risultati migliori rispetto ai precedenti: per quanto riguarda la fase di addestramento, in particolare, lo scenario risulta essere più che accettabile; infatti, l'unico valore che rende la classificazione poco ottimale è quello relativo al numero di falsi negativi che, essendo maggiore del numero di veri positivi, renderà il valore di sensibilità della rete inferiore al 50%, portando con sé gli stessi problemi precedentemente elencati.

Analizzando poi la fase di validazione della rete, si può notare come i risultati, per quanto leggermente migliori, presentino comunque le stesse problematiche della classificazione precedente; il valore riferito alla sensibilità resta anche in questo caso inferiore alla soglia minima richiesta per rendere un test diagnostico utilizzabile, il valore riferito alla specificità

migliora leggermente, mostrando così una rete più in grado di riconoscere correttamente i pazienti realmente negativi.

3.1.2 Confronto dei dati

È necessario quindi confrontare in ultimo i risultati ottenuti nei due differenti casi, cioè prima e dopo la feature selection.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.60	0.82	0.74
Validation	0.45	0.71	0.62

FS	Sensibilità	Specificità	Accuratezza
Training	0.40	1	0.78
Validation	0.40	0.77	0.64

La prima considerazione necessaria da fare è la visibile differenza di risultati che si ottengono in fase di allenamento ed in fase di validazione: infatti, i valori relativi al primo step atto alla costruzione del modello, sono in generale abbastanza validi; in particolare, prima della feature selection si notano valori perlopiù accettabili ma comunque non ottimali e non del tutto affidabili, ma lo scenario peggiora leggermente in seguito all'applicazione della selezione delle caratteristiche, successivamente alla quale il valore di sensibilità non raggiunge neppure il 50%, rendendo così il modello inadatto al riconoscimento corretto dei pazienti realmente positivi.
Tutti gli obiettivi mediocri raggiunti finora però, crollano comunque in seguito alla validazione della rete: in particolare, in entrambi i casi presentati, i modelli non sono assolutamente in grado di riconoscere e classificare correttamente i pazienti appartenenti alla classe positiva; per quanto riguarda poi i valori di specificità e di accuratezza, pur non essendo valori pessimi, comunque sia prima sia dopo l'applicazione della feature selection si nota un peggioramento dei risultati riguardante tutte le metriche di valutazione utilizzate.
In conclusione, quindi, questi modelli non sono utilizzabili in quanto, come già precedentemente detto, siano reti di classificazione relativamente funzionanti in fase di addestramento ma che poi non riescono a restituire una performance altrettanto solida e di qualità anche in fase di validazione.

3.2 Classificazione con i sintomi

Il secondo caso analizzato considera come dati di input solamente i sintomi estratti dalle cartelle cliniche dei pazienti.

A differenza dei parametri morfologici, i sintomi inclusi nello studio sono caratterizzati da una determinazione binaria della loro eventuale presenza.

Di seguito sono nuovamente raccolti i dati estratti in seguito alla media delle 100 ripetizioni effettuate sul dataset realizzato per l'allenamento delle quattro differenti reti.

RETE 1	
7.34	6.09
8.66	21.91

RETE 2	
6.87	5.92
9.13	22.08

RETE 3	
6.62	5.82
8.38	22.18

RETE 4	
4.28	3.81
11.72	24.19

Innanzitutto, è possibile notare un generale peggioramento dei risultati ottenuti dalla classificazione in cui i dati di input sono costituiti dai sintomi rispetto alle reti in cui i dati di input erano riferiti ai parametri morfologici estratti dalle immagini radiologiche, ma un confronto tra le differenti modalità sarà approfondito in seguito.

In questo caso la rete con il valore di accuratezza migliore è la prima, ma il numero di pazienti erroneamente classificati come positivi relativo a questa struttura è il maggiore ottenuto.

Nonostante questo, però sarà proprio la prima rete ad essere utilizzata: l'unico valore relativo ai falsi positivi che si distacchi di molto dagli altri è quello ottenuto in seguito all'applicazione della quarta rete; in questo caso però il valore di accuratezza risulta essere di molto inferiore rispetto a quelli ricavati dagli altri tre modelli.

Quindi, eliminando l'ultima struttura, si ottengono valori simili riguardo i pazienti erroneamente classificati per tutte le tre rimanenti reti ma si osserva anche un valore di accuratezza maggiore nella prima rete e di conseguenza un maggior equilibrio tra il numero delle due differenti classi che includono i pazienti erroneamente classificati.

Training set	
6	9
14	25

Validation set	
5	9
15	26

Il generale peggioramento osservato in fase di selezione della rete da utilizzare è purtroppo evidente anche in seguito alla sua applicazione: infatti, sia nello step di addestramento sia nella fase di validazione l'unico valore accettabile risulta essere quello relativo al corretto riconoscimento dei pazienti per i quali l'intervento non porterebbe giovamento.

In questo caso, in entrambi i passaggi, la sensibilità della rete è pessima in quanto il numero di pazienti realmente positivi ma erroneamente classificati è maggiore di quello relativo al valore

di veri positivi; inoltre, il numero di falsi positivi è decisamente troppo alto per considerare accettabile tale classificazione, in quanto, come già precedentemente specificato, il caso in cui pazienti a cui l'intervento non sia giovante vengano classificati come positivi e quindi a cui idealmente verrebbe fatto un'operazione potenzialmente rischiosa ma sicuramente inutile o addirittura peggiorativa è considerato essere lo scenario peggiore.

3.2.1 Applicazione della feature selection

In seguito all'applicazione della feature selection, i sintomi rimanenti sono quattro:

- Disturbo dei nervi cranici misti (disfonia, disfagia, singhiozzo).
- Disturbo dei nervi cranici oto-vestibolari (vertigini, acufeni).
- Disturbi sfinterici.
- Scoliosi.

RETE 1	
5.32	4.50
10.68	23.50

RETE 2	
5.38	5.11
10.62	22.89

RETE 3	
5.02	4.04
10.98	23.96

RETE 4	
3.87	3.54
12.13	24.46

Come si può facilmente osservare, in questo caso la feature selection provoca un peggioramento dei risultati già soltanto nella prima fase di analisi delle differenti reti; infatti, la feature selection è un processo molto utile per alleggerire il peso computazionale della rete ed è fondamentale per capire quali siano i parametri di ingresso che non contribuiscono in gran misura all'ottimizzazione della rete. Però a volte questo tipo di filtraggio può portare a dei problemi, come si può notare nell'esempio corrente: quando i parametri di ingresso sono pochi, questo processo può diminuire esageratamente il numero di informazioni date in pasto alla rete ed impedirle di allenarsi correttamente e di ottenere output accettabili.

Osservando i valori relativi ai falsi positivi, il dato migliore si ottiene in seguito alla quarta struttura, che però presenta un valore di accuratezza decisamente inferiore agli altri; per quanto riguarda le altre tre reti, i risultati sono molto paragonabili tra loro.

La seconda rete viene esclusa, in quanto presenti il valore relativo ai falsi positivi maggiore; osservando le due strutture rimanenti bisognerebbe selezionare la seconda ma, essendo i valori lievemente differenti tra loro e volendo mantenere un minimo di coerenza tra le strutture utilizzate, verrà scelta la prima rete come nel caso precedente.

Training set	
5	3
15	31

Validation set	
4	6
16	29

In seguito all'applicazione della feature selection la situazione presenta alcuni cambiamenti. Innanzitutto, si può notare come la fase di addestramento e la fase di validazione presentino le stesse caratteristiche e le stesse problematiche. Infatti, in entrambi i casi, si osservano valori di accuratezza accettabili e valori relativi alla specificità della rete considerabilmente buoni.
L'ostacolo resta quindi relativo alla capacità della rete di riconoscere correttamente i pazienti appartenenti alla classe positiva: infatti, in entrambe le fasi di costruzione del modello, il valore di sensibilità rimane ben al di sotto del 50%, rendendo così il modello molto in grado di gestire i pazienti per i quali l'intervento non gioverebbe ma non altrettanto capace di classificare in modo corretto i pazienti che necessiterebbero dell'operazione di decompressione.

3.2.2 Confronto dei dati

È necessario, anche in questo secondo esempio, confrontare infine i risultati ottenuti nei due differenti casi, cioè prima e dopo la feature selection.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.30	0.73	0.57
Validation	0.25	0.74	0.57

FS	Sensibilità	Specificità	Accuratezza
Training	0.25	0.91	0.67
Validation	0.20	0.83	0.60

Unendo i risultati ottenuti è possibile notare come il solo utilizzo dei sintomi presenti alla diagnosi non sia utile all'ottenimento di una classificazione robusta ed affidabile.
Prima dell'applicazione della feature selection si possono notare risultati inaccettabili sia in fase di addestramento sia in fase di validazione: i valori di sensibilità sono in entrambi i casi troppo bassi al punto da rendere totalmente falsato un ipotetico test diagnostico.
Gli unici dati leggermente accettabili sono relativi alla specificità della rete in quanto i valori di accuratezza superino di pochissimo la soglia del 50%, rendendo comunque inutilizzabile tale modello.
In seguito alla selezione delle caratteristiche la situazione migliora leggermente: sia in fase di allenamento sia in fase di validazione si può notare un netto aumento riferito ai valori di

specificità e di accuratezza, ma comunque rimane ed anzi peggiora la problematica riguardante la sensibilità e quindi la capacità di riconoscere correttamente i pazienti realmente positivi.
Comunque, anche in questo caso, nonostante l'utilizzo della feature selection apporti dei miglioramenti evidenti, entrambi i modelli presentati risultano inadeguati all'applicazione.

3.3 Classificazione con entrambi gli input

La terza casistica da analizzare è la più completa, ma ciò non significa che sia necessariamente quella che restituirà la classificazione migliore.
In quest'ultimo scenario, tutti i sintomi alla diagnosi estratti dalle cartelle cliniche e tutti i parametri morfologici ottenuti dalle immagini radiografiche sono stati uniti insieme a formare il dataset delle seguenti reti.
L'idea di inserire entrambe le tipologie di informazioni ha lo scopo di valutare l'influenza che queste abbiano sui risultati ottenuti:

- Se la prima classificazione, che includeva solamente i parametri morfologici, risultasse la migliore oppure risultasse migliore della seconda (con solamente i sintomi a comporre l'input) e paragonabile alla terza, ciò indicherebbe che i parametri estratti dalle immagini radiografiche siano sufficienti per l'ottenimento di valori accettabili e che l'aggiunta dei sintomi non dia nessun contributo al miglioramento della rete;
- Se la seconda classificazione, che includeva solamente i sintomi alla diagnosi, risultasse la migliore oppure risultasse migliore della prima (con solamente i sintomi a comporre l'input) e paragonabile alla terza, ciò indicherebbe che i sintomi estratti dalle cartelle cliniche dei pazienti siano sufficienti per l'ottenimento di valori accettabili e che l'aggiunta dei parametri non dia nessun contributo al miglioramento della rete;
- Se invece la terza classificazione risultasse la migliore, ciò vorrebbe dire che i sintomi ed i parametri morfologici debbano essere considerati insieme e che entrambe le tipologie di informazioni contribuiscano in maniera più o meno equa all'ottenimento di risultati ottimali.

RETE 1	
9.22	6.31
6.78	21.69

RETE 2	
8.25	5.17
7.75	22.83

RETE 3	
8.31	5.37
7.69	22.63

RETE 4	
7.10	4.02
8.90	23.98

Anche in questo caso i valori di accuratezza sono molto simili tra loro ma il valore relativo ai falsi positivi è nettamente migliore nella quarta rete. Nonostante, selezionando tale struttura, si favorisca un maggior numero di pazienti erroneamente classificati come negativi, in questo caso però non si può ignorare la grande differenza riguardo la condizione eletta come peggiore.

Training set	
6	0
14	34

Validation set	
4	5
16	30

In seguito all'unione di tutte le informazioni relative ai pazienti, quindi della sintomatologia e dei parametri morfologici estratti dalle immagini di risonanza, è possibile osservare alcuni fenomeni simili a quelli finora presentati

Per quanto riguarda la fase di allenamento, il valore di accuratezza raggiunge un livello abbastanza buono e di conseguenza anche il numero di pazienti erroneamente classificati diminuisce notevolmente; è però possibile notare come, nonostante il valore relativo ai falsi negativi rimanga elevato e quindi il valore relativo alla sensibilità rimarrà inferiore al 50%.

In fase di validazione poi, è purtroppo possibile notare le stesse caratteristiche osservate precedentemente: infatti il numero di pazienti realmente positivi e classificati correttamente è minore del numero di quelli classificati in modo errato, rendendo chiaro come il valore relativo alla sensibilità sarà minore del 50%. Inoltre, anche in questo caso, il numero di pazienti appartenenti alla classe dei falsi positivi, pur rimanendo comunque basso, aumenta rispetto alla fase di addestramento della rete. Queste considerazioni rendono chiaro, anche in questo caso, come il modello presentato non sia adeguato ed applicabile alla realtà.

3.3.1 Applicazione della feature selection

Anche in questo caso, per avere coerenza di risultati, è necessario applicare la feature selection; è da considerare che qui le tipologie differenti di dati di input saranno 9, cioè 5 parametri morfologici e 4 sintomi diagnostici ritenuti più rilevanti al fine della classificazione.

RETE 1	
9.42	6.30
6.58	21.70

RETE 2	
7.80	5.07
8.20	22.93

RETE 3	
7.63	4.28
8.37	23.72

RETE 4	
6.56	4.42
9.44	23.58

Si può notare come i risultati siano abbastanza simili a quelli ottenuti prima della feature selection. Anche in questo caso i valori di accuratezza sono molto paragonabili tra loro ma il dato migliore è quello relativo alla rete numero 3. In termini generali c'è un lieve peggioramento a livello dei falsi positivi rispetto alla classificazione precedente al filtraggio, però nel caso precedente il valore minore ottenuto era infatti maggiore del valore inferiore restituito dalla classificazione corrente.

In questo caso il valore minimo di falsi positivi è anch'esso da attribuire alla terza rete che, considerando anche il fatto che non sia la struttura che presenta il numero maggiore di falsi negativi, verrà selezionata per la costruzione del modello.

Training set	
11	2
9	32

Validation set	
6	13
14	22

In questo ultimo caso esposto si possono notare delle problematiche mai riscontrate finora.

In fase di allenamento i risultati ottenuti sono effettivamente validi su tutti i fronti, soprattutto in termini di accuratezza e di riconoscimento corretto della classe realmente negativa; l'unico problema è legato al numero di pazienti falsi negativi, che andrebbe oggettivamente migliorato.

Tutti gli obiettivi raggiunti in fase di allenamento vengono però demoliti in fase di validazione: non solo si osserva un valore di sensibilità inferiore al 50%, fenomeno già osservato nei modelli precedenti, ma in questo caso si nota anche un aumento inaccettabile dei pazienti classificati come falsi positivi e di conseguenza un peggioramento della rete anche nella capacità di riconoscere correttamente i pazienti realmente appartenenti alla classe negativa.

3.3.2 Confronto dei dati

In ultimo, si renderanno chiare le differenze pre e post applicazione della feature selection per osservare l'eventuale utilità del filtraggio.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.30	1	0.74
Validation	0.20	0.86	0.62

FS	Sensibilità	Specificità	Accuratezza
Training	0.55	0.94	0.80
Validation	0.30	0.63	0.52

Da queste tabelle è molto facile notare come, in questo caso, l'applicazione della feature selection non sia un processo utile ad ottenere un netto miglioramento dei risultati.
Utilizzando come dati di input sia i parametri morfologici estratti dalle immagini di risonanza magnetica sia i sintomi presenti alla diagnosi, la fase di allenamento restituisce valori quasi del tutto accettabili in entrambi i differenti modelli; nonostante il valore di specificità sia maggiore prima della feature selection, se si dovesse scegliere uno dei due modelli si dovrebbe optare per quello a cui sia stata effettivamente applicata la selezione delle caratteristiche: infatti, nonostante il valore di specificità maggiore indichi maggiore capacità della rete di riconoscere correttamente i pazienti realmente appartenenti alla classe negativa e quindi migliori risultati per quanto riguarda i falsi positivi, la sensibilità prima della feature selection risulta essere minore del 50%, valore non accettabile in quanto indichi la totale inaffidabilità del modello nella gestione dei pazienti appartenenti alla classe negativa.
Comunque, nonostante i buoni risultati ottenuti in fase di addestramento, i valori ottenuti in fase di validazione rendono inutilizzabili entrambi i modelli: non solo i valori di sensibilità restano inferiori al 50% in entrambi i casi, rendendo così le reti più in grado di fornire risultati sbagliati piuttosto che corretti, ma anche i valori di accuratezza superano di poco la metà indicando appunto la relativamente bassa probabilità che le reti siano in grado di classificare correttamente.

3.4 Confronto delle reti

Questa sezione sarà infine dedicata al confronto di tutti i risultati ottenuti e commentati fino ad ora; infatti, lo scopo primario di questa trattazione è proprio quello di valutare e capire quali siano le informazioni e le caratteristiche utili per ottenere una buona classificazione e di conseguenza un mezzo che possa aiutare i medici nella diagnosi e nella scelta del trattamento migliore per il singolo paziente.
Riassumendo, si è ritenuto opportuno lavorare tramite confronto per analizzare i risultati ottenuti; in particolare tale confronto è stato fatto tra il primo modello che vede in ingresso solamente i parametri morfologici estratti dalle immagini di risonanza magnetica per via automatica ed anche manuale, il secondo modello che prevede come dati di input soltanto i sintomi presenti alla diagnosi dei singoli pazienti ed infine il terzo modello, il cui dataset è il più completo ed è composto da tutte le informazioni disponibili riguardo i campioni analizzati.

Ad ognuna di queste strutture poi, è stata applicata la feature selection, per analizzare quali caratteristiche siano le più rilevanti e quindi quali di queste contribuiscano maggiormente all'ottenimento di un metodo di classificazione affidabile e concreto.

La prima tabella riassume quindi il primo step di questa trattazione: la costruzione dei tre differenti modelli senza però l'applicazione della feature selection. Prima di andare ad analizzare i risultati ottenuti, si potrebbe già ipotizzare che la terza struttura, che include tutte le informazioni disponibili, possa essere il metodo di classificazione di elezione, in quanto appunto sia il più completo.

Inoltre, è importante specificare che l'obiettivo sia quello di ottenere buoni risultati sia in fase di addestramento della rete sia in fase di validazione di essa: i valori estratti dai due step successivi devono essere comparabili tra loro, altrimenti la struttura non può essere considerata affidabile.

Come già specificato in precedenza infatti, non è accettabile considerare una rete che restituisca ottimi risultati in fase di training che però peggiorano in caso di validazione: questa situazione prende il nome di overfitting ed è la dimostrazione del fatto che la rete non sia in grado di svolgere il suo compito principale, cioè quello di riconoscere correttamente nuovi dati e nuovi pattern a lei sconosciuti; inoltre, non sarebbe neanche accettabile il caso opposto: se la fase di validazione risultasse ottimale e la fase di allenamento invece non restituisse valori altrettanto validi, ciò significherebbe che la rete non sia costruita bene dal principio e che probabilmente sia stata in grado di riconoscere molto bene lo specifico dataset destinato alla validazione, senza però garantire di riconoscere informazioni nuove e differenti.

Solo dopo aver fatto tutte le premesse necessarie per comprendere l'analisi dei risultati ottenuti, è possibile procedere al confronto dei differenti modelli costruiti.

SENZA FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.60	0.82	0.74
	Validation	0.45	0.71	0.62
Sintomi	Training	0.30	0.73	0.57
	Validation	0.25	0.74	0.57
Entrambi	Training	0.30	1	0.74
	Validation	0.20	0.86	0.62

Fin da subito è possibile notare come i chiarimenti fatti in precedenza fossero necessari.
Infatti, a prima vista si potrebbe dire che la prima rete, quindi quella che vede in ingresso solamente i parametri morfologici, sia decisamente la migliore: in effetti in fase di addestramento restituisce valori ottimali ed è l'unico modello a cui corrisponda un valore di sensibilità che superi la soglia del 50%; per questa ragione verrebbe spontaneo selezionarla come struttura di elezione, se non fosse però che i risultati ottenuti in seguito alla validazione siano decisamente peggiorativi. Queste caratteristiche portano quindi ad affermare che la prima rete non possa essere quella di elezione, in quanto sarebbe probabile incappare nell'overfitting e di conseguenza in un modello che non sia in grado di svolgere una classificazione affidabile ed accurata.
Nel secondo caso invece la situazione è differente: infatti, la fase di training e la fase di validazione restituiscono valori molto simili tra loro rispetto a tutte le tre metriche utilizzate per la valutazione; qui però, oltre al fatto che il problema relativo al valore relativo alla sensibilità della rete resti troppo evidente e si perpetui ad entrambe le fasi, anche i valori di accuratezza, uguali tra loro, superano di poco la soglia di accettabilità, indicando cosi che il modello non sia abbastanza in grado di classificare correttamente i campioni.
In ultimo si osserva la rete che vede come dati di ingresso tutte le informazioni disponibili insieme, quindi quelle relative ai parametri morfologici ed ai sintomi; quest'ultimo modello, ignorando per un attimo i valori relativi alla sensibilità, restituisce in effetti risultati accettabili in entrambe le fasi e decisamente migliorativi rispetto ai due scenari precedenti; questi valori, quindi, potrebbero oggettivamente essere abbastanza validi per essere presi in considerazione.
In particolare, è possibile notare gli ottimi dati ottenuti in merito alla misura di specificità di questa rete: essi dimostrano che quest'ultima struttura sia altamente in grado di riconoscere e classificare correttamente i pazienti appartenenti alla classe 0 (negativa), quindi coloro ai quali l'intervento non porterebbe giovamento. Purtroppo, però si notano anche valori nettamente inferiori rispetto alla sensibilità; la stessa struttura, quindi è decisamente non in grado di riconoscere e classificare correttamente i pazienti appartenenti alla classe 1 (positiva): ciò significa che il numero di falsi negativi, quindi quei pazienti a cui non verrebbe idealmente fatto un intervento che però sarebbe utile loro, è addirittura superiore al numero di pazienti appartenenti alla stessa classe reale ma classificati correttamente ed è quindi troppo alto per far sì che questo modello sia quello d'elezione e venga utilizzato senza remore o controlli ulteriori.

Di seguito poi vengono analizzati gli stessi modelli con però l'applicazione della feature selection: anche in questo caso è possibile fare delle previsioni sui risultati ottenuti; infatti, la selezione delle caratteristiche applicata ad un dataset già di per sé non troppo ricco, come quello contenente i sintomi presenti alla diagnosi, potrebbe portare ad una riduzione delle informazioni troppo elevata e ad una conseguente classificazione scadente.

È necessario ricordare che, in seguito all'applicazione della feature selection di tipo Neighborhood Component Analysis, le caratteristiche considerate più rilevanti sono quattro per quanto riguarda i parametri (l'area del cervello, l'area del cervelletto, l'area della fossa cranica posteriore e la lunghezza del tentorio) e quattro per i sintomi (i disturbi dei nervi cranici misti, i disturbi sfinterici, la scoliosi ed altri sintomi vari ed aspecifici).

CON FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.40	1	0.78
	Validation	0.40	0.77	0.64
Sintomi	Training	0.25	0.91	0.67
	Validation	0.20	0.83	0.60
Entrambi	Training	0.45	0.94	0.80
	Validation	0.30	0.63	0.52

Partendo dal primo modello, quindi da quello riferito solamente ai parametri morfologici, si possono già fare alcune importanti considerazioni: come è facile notare, le due fasi di costruzione della rete sono molto più equilibrate tra loro rispetto alla situazione precedente, in cui non era stata applicata alcun tipo di selezione delle caratteristiche. La rete risulta essere molto in grado di riconoscere i pazienti realmente appartenenti alla classe 0 (negativa) restituendo quindi buoni valori relativi alla misura di specificità; i valori di accuratezza sono anch'essi accettabili e si può osservare un leggero peggioramento in fase di validazione: questo fenomeno non sorprende, in quanto sia facile che una qualsiasi rete abbia più difficoltà a classificare correttamente pattern nuovi rispetto a quelli su cui è stata costruita ed addestrata. Infine, per quanto riguarda i valori di sensibilità, che indicano la capacità della rete di riconoscere i pazienti appartenenti alla classe 1 (positiva), si nota lo stesso problema osservato già nei casi precedentemente analizzati: infatti, sia la fase di addestramento sia la fase di validazione, presentano valori di sensibilità decisamente inferiori al 50%, rendendo così più

probabile che, in seguito alla classificazione dei pazienti appartenenti alla classe positiva, il risultato sia errato.

In seguito, poi, viene analizzata la rete il cui dataset è composto solamente dai sintomi diagnostici a cui è stata applicata la feature selection: nonostante le premesse, l'applicazione della feature selection permette di osservare alcuni miglioramenti: In fase di allenamento, i valori relativi alla specificità ed all'accuratezza sono superiori rispetto al caso precedente (senza l'applicazione della selezione delle caratteristiche), ma il valore relativo alla sensibilità è del tutto inappropriato. La situazione peggiora ancora in fase di validazione, con un generale abbassamento dei risultati ottenuti e con il valore di sensibilità ulteriormente peggiorato.

In ultimo verrà analizzato l'ultimo modello, quello che vede come dati di input tutte le informazioni disponibili: è facile notare come quest'ultimo scenario restituisca i valori migliori in fase di addestramento, raggiungendo ottimi risultati riguardo all'accuratezza ed alla specificità, ma l'abbassamento dei valori in fase di validazione per poter considerare tale struttura come quella di elezione.

In conclusione, selezionare il modello migliore non è facile: aver ottenuto praticamente tutti i valori riferiti alla sensibilità delle strutture inferiori al 50% rende le reti non affidabili e robuste; a questo punto però, risulterà più opportuno selezionare la rete con il valore di sensibilità inferiore. Infatti, la sensibilità è indice della capacità della rete di riconoscere correttamente i pazienti realmente appartenenti alla classe positiva; i risultati ottenuti finora quindi, indicano il fatto che le reti abbiano più probabilità si sbagliare la classificazione dei pazienti che trarrebbero giovamento dall'operazione. Scegliere la rete con la sensibilità minore vuol dire avere maggior certezza dell'errore: nella realtà significherebbe che, in caso di esito negativo del test, essendoci molta probabilità di risultati falsi negativi, questo andrebbe rifatto oppure verificato tramite altre analisi ed altri esami.

Alla luce di ciò, chiarendo che comunque non si tratti di un modello da utilizzare come gold standard, la struttura migliore ottenuta è quella che ha come input sia i parametri morfologici estratti in modo automatico dalle immagini di risonanza sia i sintomi presenti alla diagnosi, senza però l'applicazione della feature selection: questa infatti, oltre a presentare il valore relativo alla sensibilità minore rispetto agli altri modelli, risulta la rete più robusta ed affidabile in termini di accuratezza e di specificità, avendo anche una buona coerenza tra i risultati ottenuti in fase di addestramento ed i risultati ottenuti in fase di validazione.

4. Risultati del classificatore SVM

L'analisi dei risultati in seguito all'applicazione del classificatore di tipo SVM verrà svolta in modo leggermente diverso rispetto a quella fatta per la rete neurale precedentemente descritta. Infatti, non essendoci più la fase di scelta della rete ma avendo come output semplicemente i risultati della classificazione esplicitate attraverso le tre metriche finora usate, sarà sufficiente in questo caso, ma anche nel caso dei due differenti classificatori che verranno analizzati successivamente, analizzare i valori relativi alla sensibilità, alla specificità ed all'accuratezza dei modelli utilizzati. Restano però costanti i casi presi in analisi: anche per i tre classificatori osservati verranno considerati i tre differenti dataset di input, cioè in primo luogo composto solo dai parametri morfologici, successivamente composto solamente dai sintomi presenti alla diagnosi ed in ultimo composto da entrambi unitamente; infine, ad ognuno dei tre scenari verrà nuovamente applicata la feature selection.

4.1 Classificazione con i parametri morfologici

Il primo caso analizzato, come già visto per la rete neurale precedentemente trattata, è quello il cui input è composto solamente dai parametri morfologici estratti, attraverso un algoritmo automatico, dalle immagini di risonanza magnetica. I risultati relativi alla situazione antecedente ed a quella successiva all'applicazione della feature selection vengono presentati insieme per poter evidenziare, già solo alla prima occhiata, le differenze tra i due scenari.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.50	0.61	0.57
Validation	0.40	0.43	0.42

FS	Sensibilità	Specificità	Accuratezza
Training	0.62	0.39	0.48
Validation	0.60	0.54	0.56

Prima dell'applicazione della feature selection, i risultati ottenuti non sono decisamente ottimali: in fase di allenamento, infatti, tutte le tre metriche utilizzate per la valutazione risultano essere pari o di poco superiore al 50%. In particolare, il valore ottenuto di sensibilità è il peggiore che si possa ottenere: un valore esattamente pari al 50% indica la stessa probabilità che il classificatore faccia una previsione corretta oppure errata riguardo la categorizzazione dei pazienti in questo caso realmente appartenenti alla classe positiva. Tale risultato inficia del tutto il classificatore ed in confronto sarebbe paradossalmente migliore o comunque più chiaro un valore molto basso: ciò significherebbe di certo ripetere in test in caso

di paziente classificato come negativo, ma almeno si potrebbe essere certi di non potersi affidare allo strumento utilizzato. Inoltre, la fase di validazione mostra risultati ancora peggiori, rendendo chiaro il fatto che questo modello non sia assolutamente in grado di classificare correttamente gli elementi e che, di conseguenza, ci sia un'alta percentuale di errore.
In seguito all'applicazione della feature selection poi, i risultati si invertono: infatti la fase di validazione, pur mostrando valori comunque troppo bassi per considerare solida ed affidabile tale classificazione, risulta essere migliore rispetto alla fase di allenamento, in cui la classificazione risulta accettabile solo nel riconoscimento dei pazienti realmente positivi. Questo fenomeno è molto strano in quanto, teoricamente la fase di allenamento dovrebbe essere sempre migliore rispetto a quella di validazione, essendo la prima il processo di costruzione vero e proprio.

4.2 Classificazione con i sintomi

Il secondo caso analizzato è quello in cui il dataset è composto solamente dai sintomi presenti alla diagnosi; si ricorda inoltre che i sintomi presi in considerazione sono otto, tra cui disturbi cefalea nucale, disturbi dei nervi cranici misti, disturbi dei nervi cranici oto-vestibolari, disturbi delle vie lunghe, segni cerebellari, disturbi sfinterici, scoliosi ed altri sintomi aspecifici; di questi, in seguito all'applicazione della feature selection, ne verranno presi in considerazione solamente 4: disturbi dei nervi cranici misti, disturbi sfinterici, scoliosi ed altri sintomi aspecifici.

No FS	Sensibilità	Specificità	Accuratezza
Training	0	1	0.64
Validation	0	0.88	0.56

FS	Sensibilità	Specificità	Accuratezza
Training	0	1	0.64
Validation	0	1	0.64

Nel primo caso, i risultati, pur non essendo ottimali, comunque mostrano delle caratteristiche interessanti: si può notare come, in entrambe le fasi, il classificatore risulti del tutto incapace di classificare correttamente i pazienti realmente appartenenti alla classe positiva; inoltre, in fase di allenamento, il modello utilizzato è perfettamente in grado di riconoscere i pazienti realmente negativi. Questi due fattori insieme indicano il fatto che il test, almeno in fase di allenamento, risulterà sempre negativo e che, nonostante l'accuratezza raggiunga un valore limitatamente accettabile, sarà quindi inutile applicarlo. In fase di validazione poi, la situazione cambia leggermente: il fatto che il valore relativo alla specificità sia diverso dall'unità, significa

che il test in questo caso potrebbe dare risultato positivo ma che in tal caso l'output sarebbe sicuramente errato, essendo la sensibilità del classificatore pari a zero.
In seguito all'applicazione della feature selection poi, si riscontrano le stesse problematiche già analizzate nella fase di allenamento precedente, ma questa volta sono estese anche alla fase di validazione: il classificatore risulta del tutto inutile in quanto l'output sarebbe sempre negativo.

4.3 Classificazione con entrambi gli input

In ultimo verrà considerato il caso in cui i parametri morfologici estratti dalle immagini di risonanza magnetica ed i sintomi presenti alla diagnosi sono uniti a comporre il dataset.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.69	0.75	0.73
Validation	0.35	0.43	0.40

FS	Sensibilità	Specificità	Accuratezza
Training	0.56	0.78	0.70
Validation	0.30	0.91	0.69

Prima dell'applicazione della feature selection, quindi utilizzando tutte le 23 variabili disponibili, la fase di allenamento restituisce valori molto buoni, soprattutto se paragonati ai casi finora analizzati. Il comportamento del classificatore è stabile e coerente nel corretto riconoscimento dei pazienti appartenenti ad entrambe le classi. Tutti gli obiettivi raggiunti però, non risultano continuativi in fase di validazione: come è facile osservare infatti, tutte le tre metriche utilizzate per valutare la qualità della classificazione peggiorano drasticamente, rendendo il classificatore non in grado di riconoscere correttamente i pazienti e chiarendo in fatto che ci si trovi davanti ad un evidente caso di overfitting.
In seguito all'applicazione della feature selection, lo scenario presentato è molto simile a quelli incontrati durante l'analisi dei risultati ottenuti attraverso l'utilizzo della rete neurale: in fasi di allenamento i valori ottenuti sono più che accettabili, soprattutto nel corretto riconoscimento dei pazienti effettivamente negativi; in fase di validazione poi, i valori rimangono costanti o addirittura migliori (come nel caso della specificità) a meno della sensibilità: anche in questo caso il classificatore, in fase di validazione, risulta non riuscire a classificare sufficientemente bene i pazienti realmente appartenenti alla classe positiva.

4.4 Confronto dei dati

Infine, anche nel caso del classificatore SVM risulta necessario confrontare i risultati ottenuti in seguito all'utilizzo dei sei differenti dataset di input per poter valutare quale sia la struttura migliore e più adeguata ad un'ipotetica applicazione alla realtà.

SENZA FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.50	0.61	0.57
	Validation	0.40	0.43	0.42
Sintomi	Training	0	1	0.64
	Validation	0	0.88	0.56
Entrambi	Training	0.69	0.75	0.73
	Validation	0.35	0.43	0.40

Come già detto in precedenza, quando sono stati analizzati uno alla volta i differenti casi, il classificatore che aveva come input solamente i dati riferiti ai sintomi presenti alla diagnosi è da eliminare già in partenza: infatti in fase di allenamento, la sensibilità pari a zero indica che tutti i pazienti realmente positivi verranno erroneamente classificati come negativi e la specificità pari ad uno indica che tutti i pazienti realmente negativi verranno correttamente classificati come tali. Nonostante l'accuratezza sia maggiore del 50% e questo indichi che nella maggior parte dei casi l'output del classificatore sia corretto, comunque la classificazione risulterà sempre negativa e ciò rende del tutto inutile l'utilizzo di tale strumento, in quanto il valore d'uscita è noto a priori.

Con tale premessa, si potrebbe prevedere che i sintomi non diano un grande contributo al miglioramento della classificazione anche utilizzati in combinazione con i parametri morfologici.

Effettivamente, è facile notare come le differenze non siano notevoli tra il primo caso analizzato ed il terzo: in particolare, per quanto la fase di addestramento risulti migliore, indicando il fatto che l'unione delle componenti crei un dataset più adeguato alla costruzione del classificatore, la fase di validazione presenta praticamente gli stessi valori nella prima e nella terza struttura presentata, quindi hanno in comune le problematiche legate all'incapacità di classificare correttamente gli elementi nuovi.

Di seguito poi, vengono riportati i valori delle stesse strutture ma a cui è stata applicata la feature selection; visti i risultati precedenti, si può predire come la selezione delle caratteristiche è possibile che non apporti alcun miglioramento nella classificazione basata solamente sui sintomi presenti alla diagnosi.

CON FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.62	0.39	0.48
	Validation	0.60	0.54	0.56
Sintomi	Training	0	1	0.64
	Validation	0	1	0.64
Entrambi	Training	0.56	0.78	0.70
	Validation	0.30	0.91	0.69

In effetti, è facile notare come le problematiche legate alla seconda struttura persistano ed in questo caso si perpetuino anche alla fase di validazione; anche in seguito all'applicazione della feature selection il classificatore costruito solamente utilizzando i dati relativi ai sintomi presenti alla diagnosi non è in grado di garantire affidabilità e, in particolare, dimostra di non essere per nulla uno strumento utile, essendo la sua risposta sempre negativa e quindi prevedibile ancor prima dell'utilizzo.

Nella prima struttura si può notare invece un'inversione di risultati rispetto allo stesso modello senza però l'uso della feature selection: infatti, in questo caso, la fase di allenamento risulta instabile soprattutto nel corretto riconoscimento dei pazienti realmente negativi, indicando che ci sia più probabilità di incorrere nella presenza di falsi positivi piuttosto che nella comparsa di veri negativi, condizione già definita come peggiore; poi, in fase di validazione, lo scenario cambia decisamente, mostrando dei valori relativi alle metriche sostanzialmente affidabili. Questa però rimane una condizione non accettabile: per quanto, in fase di validazione, il classificatore abbia un comportamento adeguato, il fatto che la fase di addestramento non sia altrettanto di qualità indica che la costruzione del classificatore non sia andata a buon fine e che la probabilità di errore nel riconoscimento di nuovi elementi sconosciuti (quindi in fase di test del modello) sia decisamente elevata.

In ultimo si può notare come i risultati ottenuti nella terza struttura, in cui i parametri morfologici ed i sintomi sono uniti a formare il dataset, siano paragonabili a quelli ottenuti in

quasi tutte le strutture a cui era stata applicata la rete neurale: infatti, per quanto la fase di addestramento non sia la migliore, comunque mostra risultati abbastanza validi; in seguito alla validazione però, i risultati migliorano decisamente, escluso il dato relativo alla sensibilità del classificatore. Ciò indica come il classificatore non sia in grado di classificare correttamente i pazienti effettivamente positivi, restituendo più probabilmente una predizione errata a riguardo; questo, nella realtà, significherebbe che il test diagnostico andrebbe ripetuto oppure sostituito da analisi differenti in caso di output negativo.

Nonostante tale problematica a cui si può trovare una soluzione e nonostante non siano risultati ottimali, anche in questo caso la classificazione migliore avviene in seguito all'unione dei parametri morfologici e dei sintomi a cui però viene applicata la feature selection, essendo di fatto l'unica ad avere un certo equilibrio e valori sufficientemente validi in entrambe le fasi.

5. Risultati della KNN

L'analisi dei risultati ottenuti in seguito all'utilizzo del classificatore di tipo K-Nearest Neighbors, dove, in questo caso, come già precedentemente specificato, il valore di K è pari a 2, verrà svolta in maniera leggermente diversa, vista la particolarità dei valori.

In particolare, è possibile notare come, sia nel primo caso in cui vengono analizzati solamente i parametri morfologici sia nel terzo caso in cui vengono analizzati i parametri morfologici insieme ai sintomi, i risultati siano esattamente gli stessi; tale uguaglianza vale per entrambe le fasi di costruzione del modello ed inoltre è osservabile sia prima che dopo l'applicazione della feature selection.

Queste caratteristiche sono probabilmente dovute al fatto che tale classificazione dipenda solamente da fattori geometrici; quindi, viene considerata la distanza tra gli elementi del dataset per determinare con quale altro campione confrontare il paziente in questione.

Nonostante questo sia limitante in termini di osservazione dei fenomeni selezionati, rende comunque possibile l'avanzamento di alcune ipotesi; nello specifico, dato il fatto che la classificazione svolta solamente analizzando i parametri morfologici e quella svolta analizzando i parametri morfologici unitamente ai sintomi presenti alla diagnosi diano esattamente gli stessi risultati, sia in fase di allenamento sia in fase di validazione, è indice del fatto che i sintomi in questo caso non diano alcuna informazione e non risultino affatto utili nel miglioramento della classificazione già ottenuta grazie ai soli parametri morfologici.

5.1 Classificazione con i parametri morfologici e con entrambi gli input

In questo caso quindi, una sola tabella sarà sufficiente per esporre i valori ottenuti attraverso l'utilizzo di quattro differenti modelli; in particolare, tali risultati sono validi per la classificazione svolta tramite l'analisi dei soli parametri morfologici (prima e dopo l'utilizzo della feature selection) e per il modello in cui il dataset di ingresso è composto da tutte le informazioni disponibili, quindi quelle relative ai parametri morfologici e quelle relative ai sintomi presenti alla diagnosi, anche questa volta sia prima sia dopo l'applicazione della selezione delle caratteristiche.

	Sensibilità	Specificità	Accuratezza
Training	1	0.53	0.70
Validation	0.61	0.76	0.71

La tabella che quindi riassume tutti i modelli precedentemente elencati mostra però risultati sufficientemente validi, sia in fase di allenamento sia in fase di validazione, presentando anche un buon livello di coerenza tra i due step di costruzione del classificatore.

In particolare, contrariamente a ciò che si è osservato nei casi precedentemente analizzati, in fase di allenamento si nota un perfetto valore di sensibilità che, come spesso accade, porta ad ottenere un valore leggermente inferiore alla media riguardo la capacità del modello di riconoscere correttamente i pazienti realmente negativi; il valore di accuratezza invece raggiunge un livello accettabile che è mantenuto costante anche nella fase successiva.

In fase di validazione poi, l'andamento è invertito e più simile a ciò che si è potuto notare finora: il valore di specificità cresce ed il modello ha maggiori difficoltà nel riconoscere correttamente i pazienti effettivamente positivi.

5.2 Classificazione con i sintomi

L'unico caso ad avere risultati diversi da quelli precedentemente analizzati è quello in cui il dataset di ingresso è composto solamente dai sintomi presenti alla diagnosi.

Prima della feature selection è immediato notare come tale classificazione non sia assolutamente utilizzabile e si ripresenta un ostacolo già precedentemente osservato: infatti, un valore di sensibilità così alto indica che tutti i pazienti realmente positivi verranno classificati

come tali; inoltre, un valore di specificità altrettanto basso indica l'altissima probabilità che i pazienti realmente negativi vengano invece classificati come positivi. Ciò significa che, non solo il test non sarebbe affidabile visti i risultati ottenuti, ma non sarebbe neppure utile considerando che il valore di output positivo sia prevedibile e praticamente certo.

No FS	Sensibilità	Specificità	Accuratezza
Training	1	0.07	0.41
Validation	0.45	0.77	0.58

FS	Sensibilità	Specificità	Accuratezza
Training	1	0.18	0.47
Validation	0.41	0.89	0.49

In seguito all'applicazione della feature selection poi, si nota un leggero miglioramento in termini di specificità e quindi di accuratezza, ma comunque non sufficiente da far sì che questo modello possa essere preso in considerazione. Inoltre, è possibile notare un'inversione di scenario in seguito alla validazione; infatti, in questo caso, l'unico valore accettabile ed anzi molto alto è proprio quello relativo alla specificità, ma la capacità di riconoscere correttamente i pazienti realmente negativi è l'unica caratteristica valida di quest'ultimo classificatore.

5.3 Confronto dei dati

Vista la particolarità dei risultati, è sufficiente una sola tabella per riassumere tutti i risultati ottenuti; in particolare, le prime due righe sono riferite ai valori ottenuti in fase di addestramento e di validazione dei modelli che prendono in esame come dati di input:

- Solamente i parametri morfologici estratti dalle immagini di risonanza magnetica.
- Solamente i parametri morfologici a cui viene applicata la feature selection.
- I parametri morfologici insieme ai sintomi presenti alla diagnosi.
- I parametri morfologici insieme ai sintomi a cui viene applicata la feature selection.

RISULTATI OTTENUTI		Sensibilità	Specificità	Accuratezza
Parametri ed entrambi	Training	1	0.53	0.70
	Validation	0.61	0.76	0.71
Sintomi senza FS	Training	1	0.07	0.41
	Validation	0.45	0.77	0.58
Sintomi con FS	Training	1	0.18	0.47
	Validation	0.41	0.89	0.49

Iniziando dai due modelli che prendono in esame solamente i sintomi presenti alla diagnosi, anche con l'utilizzo di questo classificatore si può affermare con certezza che queste informazioni da sole non siano sufficienti alla costruzione di un modello di classificazione robusto ed affidabile. In particolare, in questo caso, il modello, facendo riferimento alla fase di addestramento, risulterebbe persino inutile, in quanto gli alti valori di sensibilità ed i bassi valori di specificità restituirebbero un test diagnostico il cui risultato sarebbe molto probabilmente sempre positivo, prevedibile e che quindi perderebbe ogni significato che un test diagnostico debba e possa avere. Nonostante, in fase di validazione i risultati siano più equilibrati, restano comunque insufficienti per l'utilizzo di tale modello.

Per quanto riguarda poi la seconda classificazione, quindi quella che racchiude i quattro differenti modelli precedentemente elencati, lo scenario cambia radicalmente mostrando un netto miglioramento; tutti i valori relativi alle metriche di valutazione utilizzate risultano sufficienti, mostrando una particolarità in termini di sensibilità: infatti, in fase di allenamento, tale caratteristica mostra un valore unitario, diversamente da tutti i casi finora utilizzati, in cui il problema principale era proprio legato alla corretta classificazione dei pazienti realmente positivi. Inoltre, è importante notare come i valori delle metriche in fase di validazione siano più equilibrati tra loro; questo è un punto fondamentale per qualsiasi classificatore, in cui appunto di cerca un corretto compromesso tra il valore di sensibilità e quello di specificità.

Infine, è necessario selezionare, anche in questo caso, il modello risultato migliore e più performante: la scelta ovviamente ricade sui risultati che racchiudono i quattro differenti dataset di ingresso; tra questi quattro modelli però, si selezionerà quello che analizza solamente le informazioni riguardanti i parametri morfologici a cui è stata applicata la feature selection. Infatti, per quanto concerne la scelta delle features da considerare, è ovvio che, a parità di risultati, sia meglio selezionare il dataset più leggero, osservando che le informazioni aggiuntive (in questo caso relative ai sintomi presenti alla diagnosi) non diano alcun contributo nel miglioramento della classificazione. L'ultima questione, riferita alla feature selection, segue lo stesso ragionamento: a parità di risultati, è sempre meglio ridurre il peso computazionale ed il tempo di elaborazione dei dati, preferendo quindi il dataset privato delle informazioni superflue oppure ridondanti.

6. Risultati di Naive-Bayes

Come già spiegato in precedenza, quest'ultimo classificatore utilizzato si basa sulla probabilità, calcolata attraverso il teorema di Bayes appunto, che un dato elemento di ingresso appartenga ad una delle due classi definite. La denominazione di tipo "naive" deriva poi dal fatto che le ipotesi di partenza siano molto semplici; infatti, queste sostengono che non ci sia interdipendenza tra le caratteristiche e che queste debbano essere considerate singolarmente. Uno dei principali vantaggi di questo strumento è che sia in grado di funzionare molto bene anche nel caso in cui il dataset di allenamento sia ridotto, come accade in questa trattazione. Infatti, questo classificatore è semplicemente parametrizzato attraverso la media e la varianza di ogni variabile considerata indipendente; altri strumenti di classificazione invece basano il calcolo delle probabilità anche sulla matrice di covarianza, che in caso di dataset ridotto può rappresentare valori variabili che andrebbero poi ad inficiare le prestazioni del modello.

6.1 Classificazione con i parametri morfologici

Iniziando come sempre dal modello costruito solamente utilizzando i dati relativi ai parametri morfologici estratti dalle immagini di risonanza magnetica, si possono immediatamente notare dei miglioramenti.

La prima struttura presentata mostra tutti i valori relativi alle metriche decisamente sufficienti sia in fase di allenamento sia in fase di validazione, indicando quindi una certa coerenza ed una certa robustezza. I valori più bassi sono quelli relativi alla sensibilità del classificatore, cioè legati alla capacità di esso di riconoscere e classificare correttamente tutti i pazienti positivi, ai quali l'intervento porterebbe giovamento; si ripete quindi la problematica già osservata in seguito all'applicazione delle reti neurali, ma in questo caso, nonostante il valore non sia ottimale, la probabilità di classificazione corretta è maggiore di quella di errore.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.56	0.93	0.79
Validation	0.55	0.93	0.73

FS	Sensibilità	Specificità	Accuratezza
Training	0.44	0.93	0.75
Validation	0.50	0.89	0.74

In seguito all'applicazione della feature selection però, i risultati peggiorano leggermente sia in fase di allenamento che in fase di validazione; in questo caso le problematiche legate al valore di sensibilità risultano accentuate rispetto al caso appena descritto. Infatti, nella prima fase la

probabilità di errore risulta essere maggiore della probabilità di classificare correttamente i pazienti realmente positivi; inoltre, in fase di validazione, si nota un valore di sensibilità esattamente pari al 50%: questo dato è il peggiore che si possa ottenere, in quanto non lasci spazio a previsioni affidabili. Infatti, in caso di metriche di molto maggiori del 50%, la classificazione risulta affidabile e non è necessario ripetere il test oppure integrarlo con altre analisi; in caso di metriche di molto inferiori al 50% invece, si può affermare che il test non sia affidabile e che, nel momento in cui si ottiene un determinato output (in questo caso negativo) il test sia assolutamente da ripetere perché la probabilità di errore è notoriamente alta.

6.2 Classificazione con i sintomi

Considerati i risultati ottenuti con gli strumenti precedenti in cui venivano utilizzati i soli sintomi per la costruzione della classificazione, i valori che si possono osservare in questo caso sono decisamente migliorativi.

Prima dell'applicazione della feature selection, la fase di allenamento restituisce effettivamente valori abbastanza affidabili e comunque con una maggiore probabilità di classificazione corretta per tutte le tre metriche. Purtroppo, però la fase di validazione non risulta coerente e restituisce valori completamente inutilizzabili, andando a confermare ciò che era già stato osservato in precedenza: l'utilizzo dei soli sintomi presenti alla diagnosi non è sufficiente ed adeguato all'ottenimento di una classificazione applicabile al caso reale.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.69	0.57	0.61
Validation	0.30	0.48	0.42

FS	Sensibilità	Specificità	Accuratezza
Training	0.75	0.39	0.52
Validation	0.20	0.37	0.31

In seguito all'applicazione della feature selection poi, i risultati peggiorano ulteriormente. La fase di allenamento mostra valori inferiori (fatta eccezione per la sensibilità del classificatore) e addirittura minori del 50%; in fase di validazione poi si nota lo stesso fenomeno precedentemente osservato, quindi un netto abbassamento delle metriche rendendo così inutilizzabile il classificatore. Il fatto che la feature selection potesse apportare dei peggioramenti era relativamente prevedibile sia osservando i risultati precedentemente ottenuti in seguito all'utilizzo dei soli sintomi per la classificazione sia osservando che la selezione delle caratteristiche serva proprio ad eliminare i dati ridondanti, che non contribuiscono al miglioramento della classificazione ma che aumentano solamente il peso

computazionale; quindi, l'utilizzo della feature selection su un dataset già di per sé abbastanza limitato rischia di peggiorare i risultati andando ad eliminare informazione che avrebbero anche minimamente contribuito all'ottenimento di risultati validi e robusti.

6.3 Classificazione con entrambi gli input

In ultimo vengono analizzati i risultati ottenuti in seguito all'utilizzo del dataset completo; quindi, composto sia dai parametri morfologici sia dai sintomi presenti alla diagnosi.

Prima dell'applicazione della feature selection, si possono osservare valori abbastanza validi; in particolare, in fase di allenamento, i valori relativi alla specificità ed all'accuratezza raggiungono livelli effettivamente elevati. Per quanto riguarda poi la sensibilità, quindi la capacità del classificatore di riconoscere correttamente i pazienti effettivamente appartenenti alla classe positiva, il valore risulta essere leggermente più basso, fenomeno già frequentemente osservato nei casi precedenti, ma comunque anch'esso accettabile. La situazione poi è perpetuata anche in fase di validazione del classificatore, che presenta risultati validi riguardo la capacità di riconoscere i pazienti a cui l'intervento non porterebbe giovamento e più in generale riguardo la capacità di classificare correttamente gli elementi di input. Anche in questa fase è inoltre possibile notare come il risultato riferito alla sensibilità sia inferiore, ed in particolare pari al 50%: tale valore, come già spiegato in precedenza, porta con sé dei problemi legati all'affidabilità ed alla robustezza della classificazione in esame.

No FS	Sensibilità	Specificità	Accuratezza
Training	0.62	0.96	0.84
Validation	0.50	0.71	0.64

FS	Sensibilità	Specificità	Accuratezza
Training	0.50	0.93	0.77
Validation	0.50	0.80	0.69

In seguito all'utilizzo della feature selection poi, soffermandosi inizialmente solo i valori relativi alla sensibilità ed all'accuratezza del classificatore, lo scenario rimane più o meno costante, con una diminuzione per quanto riguarda la capacità generale di riconoscere gli elementi; in fase di validazione poi, i valori restano abbastanza costanti e coerenti con quelli appena analizzati, mostrando una capacità migliore, rispetto al caso precedente in cui venivano analizzate tutte le informazioni disponibili riguardo i pazienti, nel riconoscere correttamente campioni di input nuovi e mai visti, su cui il classificatore non è stato allenato. Anche in questo caso però persiste il problema legato alla capacità del modello di riconoscere correttamente i pazienti realmente

positivi: questa struttura presenta lo stesso ostacolo già precedentemente descritto come invalidante, ed è inoltre perpetuato anche in fase di validazione del classificatore in esame.

6.4 Confronto dei dati

Anche nel caso di questo ultimo classificatore utilizzato è necessario confrontare tra loro tutti i risultati ottenuti in seguito all'applicazione dei sei differenti modelli per poter scegliere tramite quale si ottenga la classificazione più affidabile e più robusta.

SENZA FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.56	0.93	0.79
	Validation	0.55	0.93	0.73
Sintomi	Training	0.69	0.57	0.61
	Validation	0.30	0.48	0.42
Entrambi	Training	0.62	0.96	0.84
	Validation	0.50	0.71	0.64

Il primo modello utilizzato è quello relativo ai risultati ottenuti in seguito all'analisi del dataset composto solamente dai parametri morfologici estratti dalle immagini di risonanza magnetica. Questa struttura, confrontata con le altre ottenute tramite questo classificatore ma anche con quelle ottenute in seguito all'utilizzo di altri differenti metodi di classificazione, è l'unica a presentare tutti i valori relativi alle tre metriche di valutazione superiori al 50%, sia in fase di addestramento sia in fase di validazione. I risultati peggiori di questa struttura sono come sempre relativi alla sensibilità, quindi alla capacità del classificatore di riconoscere correttamente i pazienti realmente appartenenti alla classe positiva. Dato il fatto che questa tipologia di problema si perpetua tra i vari modelli analizzati ed anche tra i differenti metodi di classificazione, è chiaro come sia probabile che la questione sia relativa alla qualità dei dati riferiti a questa categoria di pazienti piuttosto che ai classificatori in sé.

Il secondo caso è relativo all'utilizzo dei soli sintomi presenti alla diagnosi per la costruzione del classificatore; come già visto in precedenza, questo dataset non risulta essere adatto all'ottenimento di valori validi e quindi all'ottenimento di una classificazione robusta. Nonostante i valori ottenuti in fase di addestramento, per quanto non siano auspicabili, siano comunque accettabili, tutta l'affidabilità del classificatore viene persa in fase di validazione, in

cui nessuna delle tre metriche di valutazione indichi una qualche capacità di riconoscimento corretto dei pazienti inclusi nello studio.

Per quanto tale classificazione non sia affidabile, l'utilizzo dei sintomi unitamente all'utilizzo dei parametri morfologici, per quanto riguarda la fase di addestramento, è in grado di fornire risultati molto validi e persino migliori rispetto al primo modello utilizzato. In fase di validazione però, i valori peggiorano leggermente e restano comunque inferiori ai dati ottenuti nella medesima fase ma utilizzando solamente i parametri morfologici. Per questo motivo, ma anche per il fatto che il valore di sensibilità riferito a quest'ultimo modello sia pari al 50%, fino ad ora la struttura migliore rimane la prima analizzata, contrariamente a quanto osservato precedentemente, in cui la classificazione migliore si otteneva in seguito all'utilizzo dei sintomi presenti alla diagnosi ed ai parametri morfologici insieme.

CON FEATURE SELECTION		Sensibilità	Specificità	Accuratezza
Parametri morfologici	Training	0.44	0.93	0.75
	Validation	0.50	0.89	0.74
Sintomi	Training	0.75	0.39	0.52
	Validation	0.20	0.37	0.31
Entrambi	Training	0.50	0.93	0.77
	Validation	0.50	0.80	0.69

In seguito all'applicazione della feature selection poi, i risultati cambiano decisamente.

Nel prima caso, infatti, la fase di allenamento peggiora decisamente, raggiungendo addirittura valori minori del 50% in termini di sensibilità; in fase di validazione poi, il risultato è abbastanza coerente ed anche migliorativo nel riconoscimento dei pazienti positivi. Queste caratteristiche non rendono comunque sufficientemente affidabile, al punto di essere considerata migliore della medesima struttura a cui però non era stata applicata la feature selection.

Nel secondo caso poi, la selezione delle caratteristiche peggiora ulteriormente la classificazione ottenuta, rendendo definitivamente chiaro il fatto che i sintomi presenti alla diagnosi non portino informazioni sufficienti per la creazione di un test diagnostico applicabile alla realtà. Come già detto in precedenza, il fatto che la feature selection potesse essere un'operazione peggiorativa nel caso della classificazione basata solamente sulla sintomatologia era

prevedibile in quanto le informazioni ad essa relative non erano ridondanti o superflue già in partenza.

In ultimo, i risultati relativi al terzo modello presentato non sono molto diversi rispetto a quelli ottenuti con la stessa struttura ma senza l'applicazione della feature selection e non risultano neanche essere i migliori tra tutti i tre casi a cui è stata applicata la selezione delle caratteristiche. Anche a quest'ultimo scenario possono essere attribuite le stesse problematiche già riscontrate in precedenza: nonostante i valori relativi alla specificità ed all'accuratezza siano validi e coerenti in entrambe le fasi di costruzione del modello, la sensibilità è nuovamente pari al 50%, indicando inaffidabilità e robustezza insufficiente.

Anche in seguito alla feature selection quindi, il modello migliore risulta essere quello relativo all'utilizzo delle informazioni estratte solamente dai parametri morfologici; nonostante ciò, il classificatore di Naive-Bayes migliore risulta essere il primo, quello in cui non veniva applicata la feature selection al dataset composto dai parametri morfologici, in quanto questo sia l'unico a presentare valori superiori al 50% per tutte le tre metriche di valutazione utilizzate per entrambe le fasi di costruzione svolte.

7. Confronto dei metodi di classificazione

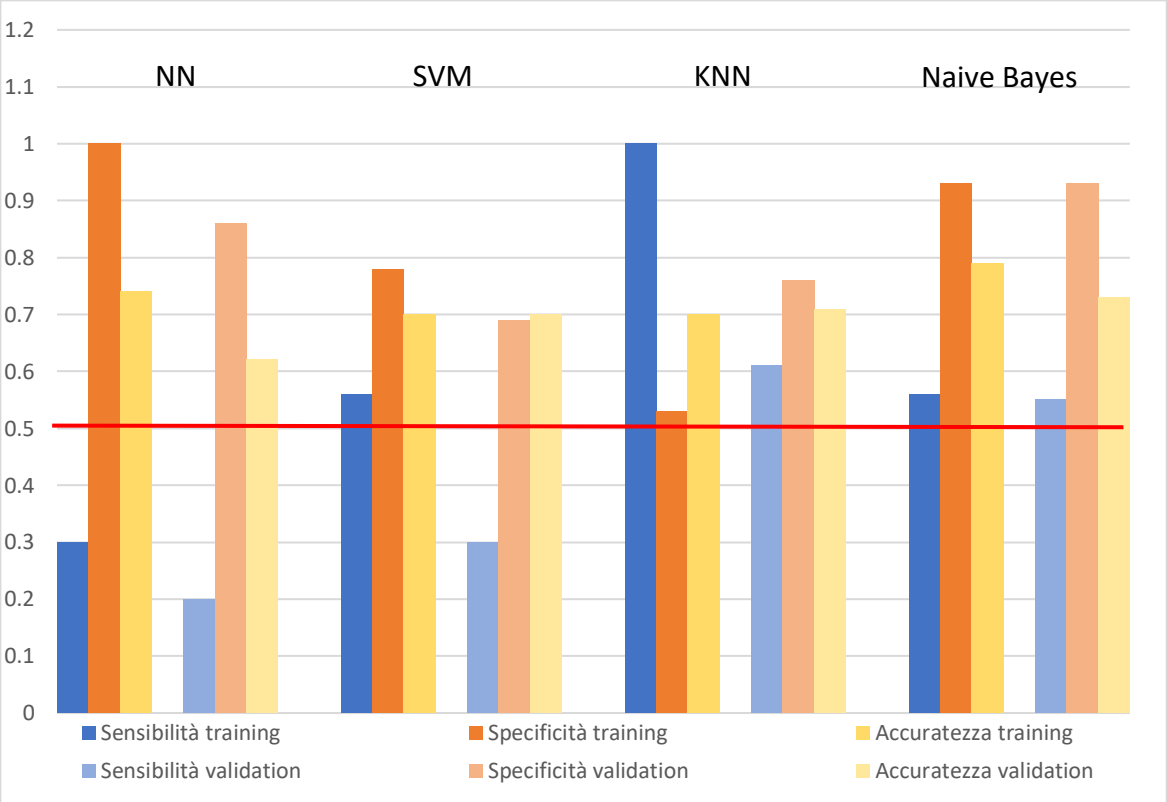
L'obiettivo di questa trattazione è l'analisi dell'influenza che differenti fattori possano avere nell'ottenimento di una classificazione robusta ed affidabile; in particolare sono state considerate le informazioni derivanti dai quindici parametri morfologici estratti dalle immagini di risonanza magnetica, le informazioni relative agli otto sintomi clinici presenti alla diagnosi ed in ultimo sono state analizzate le caratteristiche più importanti tra quelle appena elencate.

Per fare ciò sono stati analizzati sei differenti modelli a cui sono stati applicati quattro metodi di classificazione; per ognuno di questi metodi poi, è stato scelto il modello più performante sulla base delle tre metriche di valutazione selezionate: la sensibilità, la specificità e l'accuratezza. Infine, è necessario quindi selezionare quale classificatore dia i risultati migliori, sia per determinare quale sia il classificatore più performante sia per comprendere quali siano le caratteristiche più rilevanti al fine di una corretta classificazione.

Prima di giungere al confronto vero e proprio, è necessario ricapitolare quali modelli siano stati scelti per i differenti classificatori:

- Nel primo caso la classificazione viene svolta utilizzando un metodo di machine learning, quindi una rete neurale semplice (di tipo feedforward) e il dataset che restituisce i risultati migliori è quello composto da entrambe le tipologie di informazioni, quindi dai parametri morfologici e dai sintomi, a cui non viene però applicata la feature selection.
- Nel secondo caso lo strumento utilizzato è un classificatore SVM ed il modello migliore scaturisce dall'utilizzo del dataset completo, quindi lo stesso del caso precedente, a cui però questa volta viene applicata la feature selection.
- Nel terzo caso il classificatore è di tipo KNN, strumento che restituisce risultati abbastanza particolari, come già chiarito in precedenza; questa volta il modello più performante è ottenuto in seguito all'utilizzo di un dataset composto solamente dai parametri morfologici a cui viene applicata la feature selection.
- Nell'ultimo caso il classificatore è di tipo Naive-Bayes e, con tale strumento, i risultati migliori si osservano conseguentemente all'utilizzo di un dataset composto solamente dai parametri morfologici a cui però non viene applicata feature selection.

In seguito a tali selezioni, è stato possibile costruire il grafico sottostante, che permette di confrontare chiaramente i risultati ottenuti dai quattro modelli migliori, sulla base delle tre metriche di valutazione scelte e riferiti ad entrambe le fasi di costruzione di un classificatore.



Per ogni metodo di classificazione sono quindi riportati sei valori: i primi tre riferiti alle tre metriche di valutazione riguardanti la prima fase di addestramento e i restanti relativi ai medesimi tre parametri ma inerenti alla fase di validazione.

Oltre ai valori di sensibilità, specificità ed accuratezza, è riportato anche un valore di soglia pari al 50%: nonostante il superamento di tale valore non sia sufficiente per considerare affidabile una qualsivoglia classificazione, questo è il minimo obiettivo da raggiungere per poter prendere in considerazione un modello, in quanto un valore inferiore indicherebbe una maggior probabilità di errore piuttosto che di ottenere una classificazione corretta.

Il primo modello presentato è appunto quello costruito attraverso l'utilizzo di una rete neurale molto semplice che prende in ingresso tutte le informazioni disponibili sui pazienti, quindi i parametri morfologici ed i sintomi senza l'applicazione della feature selection. È facile notare come, in questo modello, ci sia un chiaro problema legato alla sensibilità della rete, quindi alla sua capacità di riconoscere correttamente i pazienti effettivamente positivi; questa problematica è stata osservata svariate volte nel corso della trattazione, ed in particolare è stata osservata in tutti i sei modelli a cui è stata applicata la rete neurale. Per questo motivo, è stata quindi scelto il dataset di ingresso con buoni valori rispetto alle altre due metriche ma valori peggiori proprio rispetto alla sensibilità: così facendo, perlomeno si può avere una certezza di errore maggiore, andando così a ripetere il test in caso di risultato negativo. Inoltre, si nota come i valori riferiti alla fase di validazione siano quasi proporzionalmente minori di quelli relativi alla fase di addestramento, indicando così un piccolo ostacolo nel riconoscimento di elementi nuovi su cui la rete non è stata addestrata. Questi due fattori quindi, fanno sì che questo non sia il modello di elezione, ma nonostante ciò la soluzione della rete neurale non andrebbe eliminata ma piuttosto approfondita e migliorata; per ottenere risultati migliori infatti, sarebbe bene fare un controllo a ritroso sui pazienti positivi e su tutte le informazioni derivanti da essi, soprattutto quelle riguardanti i parametri morfologici: infatti, l'errore potrebbe essere proprio sul modo, ad esempio sull'algoritmo automatico, in cui tali parametri geometrici vengono estratti dalle immagini di risonanza magnetica.

Il secondo modello è ottenuto in seguito all'utilizzo di un classificatore di tipo SVM, in cui si analizzano solamente gli elementi marginali, cioè più vicini alla linea di separazione tra le due classi; anche in questo caso il dataset è composto sia dai parametri morfologici sia dai sintomi. In questo caso, il problema legato alla sensibilità si risolve, perlomeno in fase di addestramento, ma di conseguenza si nota anche una diminuzione in termini di specificità ma anche di

accuratezza; sicuramente in questo modello il compromesso tra le due metriche opposte è migliore rispetto a quello ottenuto con la struttura precedente. Nonostante ciò, in fase di validazione la sensibilità del classificatore torna ad essere decisamente bassa ed inferiore al limite di accettabilità, mostrando quindi incoerenza tra le due fasi, fenomeno che rende il modello inutilizzabile.

Nel caso del classificatore k-NN il dataset di elezione è composto solamente dai parametri morfologici a cui è stata applicata la feature selection; è però necessario ricordare che i medesimi risultati fossero stati ottenuti anche in seguito all'analisi dei parametri e dei sintomi presenti alla diagnosi insieme. È immediato notare come questa classificazione sia la più adeguata finora, e per tale affermazione è sufficiente osservare come, per la prima volta, tutte le tre metriche, sia per la fase di addestramento sia per la fase di validazione, siano superiori alla soglia minima imposta. Inoltre, è possibile apprezzare anche un certo livello di coerenza tra le due fasi di costruzione del classificatore, fatta eccezione solamente per il valore di sensibilità.

L'ultimo scenario è quello ottenuto in seguito all'utilizzo del classificatore di tipo Naive-Bayes, in cui tutte le singole caratteristiche sono considerate come indipendenti dalle altre; in questo caso, infine, il dataset di elezione è composto solamente dai parametri morfologici a cui non è stata applicata la feature selection. Il classificatore in questione è anch'esso in grado di restituire come output dei valori riferiti alle metriche di valutazione tutte superiori al valore di soglia, sia in fase di addestramento sia in fase di validazione. Nonostante tale caratteristica sia comune anche al modello k-NN precedentemente utilizzato, il modello Naive-Bayes sarà proprio quello di elezione, e questa scelta è dettata proprio da uno dei criteri selezionati all'inizio della trattazione: infatti, una delle differenze principali tra gli ultimi due classificatori utilizzati è legata al valore di specificità; nel modello di Naive-Bayes, nonostante la sensibilità sia di molto inferiore rispetto a quello ottenuto con il k-NN, la capacità di riconoscere correttamente gli elementi negativi è migliore e di conseguenza sarà inferiore il numero di falsi positivi, quindi di pazienti che incorrerebbero in un intervento non giovante per errore del test diagnostico.

8. Conclusioni

La malformazione di Chiari, ad oggi, non ha ancora raggiunto un iter standard per la cura ed il corretto decorso chirurgico: finora è stato ritenuto sufficiente basarsi su una soglia minima di

discesa tonsillare per diagnosticare tale patologia ed è stato scelto l'intervento di decompressione osteo-legamentosa come gold standard riguardo il trattamento da seguire. Questo tipo di operazione, infatti, essendo relativamente semplice e sicura, ed in molti casi anche utile e risolutiva, è stata scelta come quella principale in caso di diagnosi della malformazione di Chiari.

Come risulta però chiaro dagli studi condotti dalla facoltà di Medicina e Chirurgia dell'Università di Torino, purtroppo l'intervento di decompressione non può essere selezionato come primo ed unico intervento risolutivo del problema: infatti, tale studio ha seguito un follow up fino a 5 anni dalla diagnosi, riportando tutti i dati relativi allo sviluppo di problemi post operatori (ad esempio la siringomielia), ai casi in cui l'intervento di decompressione sia stato rapidamente seguito da un'altra differente operazione più invasiva e quindi in cui i pazienti siano stati sottoposti a rischi legati ad una qualsiasi pratica chirurgica che si sarebbero potuti evitare oppure ai casi in cui l'intervento non abbia apportato alcun miglioramento.

Visti tutti i problemi fino ad ora elencati, lo scopo principale di questo progetto è proprio quello di eliminare questi ostacoli e provare a trovare un percorso da seguire sicuro e che portasse solamente giovamento a tutti i pazienti affetti da malformazione di Chiari; per fare ciò, dal punto di vista ingegneristico si è scelto, come ormai è noto, di utilizzare il mezzo più utile in questi casi: il machine learning.

Ovviamente l'obiettivo va raggiunto passo dopo passo, e questa trattazione, pur non essendo risolutiva, ha aggiunto informazioni utili per il proseguimento delle ricerche.

Infatti, fino a prima, si era cercato di capire ed analizzare quali fossero i parametri morfologici fondamentali per diagnosticare la malformazione di Chiari e per determinare quale tipo di intervento sarebbe stato il più adeguato al singolo paziente.

Ovviamente i parametri estraibili da un'immagine di risonanza magnetica sono infiniti e non sono stati considerati tutti: ciò nonostante, dagli studi precedenti, sono emerse informazioni importanti riguardo la rilevanza dei differenti parametri, ma comunque non sufficienti e solide al punto da poter basare l'intero percorso di cura solamente su di esse.

Questi risultati incerti possono essere stati causati da differenti fattori: nel momento in cui si utilizzano delle reti neurali o dei classificatori per ricerca bisogna essere in grado di selezionare la struttura più adeguata allo scopo che si vuole raggiungere ed ai dati che si hanno a disposizione; inoltre, i parametri utilizzati sia in questo caso che nel caso precedente, sono stati estratti automaticamente attraverso l'utilizzo di un algoritmo: questo porta con se tutta una

serie di problemi che possono essere legati alla struttura dell'algoritmo in sé, alla qualità delle immagini che, per quanto possa essere ottimizzata ed uniformata, non potrà mai essere del tutto standardizzata. Inoltre, un grande limite di questo progetto in generale, è legato al numero di immagini disponibili: essendo un caso sperimentale, per includere i pazienti allo studio è necessaria la loro autorizzazione ed è necessario seguire l'iter etico che garantisca sicurezza e privacy a tutti i partecipanti; in aggiunta, essendo forte la volontà di portare avanti lo studio sia dal punto di vista ingegneristico sia dal punto di vista medico e di garantire continuità e coerenza tra le due parti differenti ma complementari, è stato necessario giungere a compromessi tra i criteri di inclusione selezionati dai medici e quelli selezionati dagli ingegneri.

Così facendo, però, in seguito a questa selezione sono rimasti solamente 55 pazienti utilizzabili: questo numero non è troppo basso per poter ottenere dei risultati validi, ma per quanto questi possano essere anche molto buoni, in futuro andrebbero sicuramente testati su un pool molto più esteso di campioni.

Tornando quindi a questa specifica trattazione, l'idea è stata quella di non soffermarsi solamente sui parametri morfologici e sulla loro influenza sul decorso della patologia e dell'iter chirurgico, evitando di provare a risolvere tutti i problemi computazionali che potevano essere presenti in seguito all'utilizzo di un algoritmo automatico ma spostando invece il focus su altre informazioni: i sintomi presenti alla diagnosi manifestati da ogni paziente incluso nello studio.

Questa scelta è stata dettata dalla volontà di mantenere una coerenza con lo studio portato avanti dall'Università degli studi di Torino ma anche di non considerare solamente meri dettagli geometrici estraibili dalle immagini di risonanza magnetica, ma di analizzare la patologia in questione in toto, quindi analizzando anche tutte le altre informazioni disponibili sullo stato clinico del singolo paziente sottoposto a trial clinico e chirurgico.

Tale analisi è stata quindi portata avanti semplicemente per confronto: si sono analizzate le singole influenze dei parametri morfologici e dei sintomi separatamente, per poi unire la loro rilevanza utilizzandoli entrambi come input delle reti neurali.

Si giunge ora alla selezione del modello di classificazione che restituisca i risultati migliori: come già visto in precedenza, il test diagnostico idealmente migliore si otterrebbe utilizzando il classificatore Bayesiano semplificato, il dataset composto solamente dai parametri morfologici e senza l'utilizzo della feature selection; nonostante questa struttura risulti la più performante,

è necessario osservare in toto i risultati ottenuti per comprendere meglio l'andamento della trattazione.

Partendo dall'ipotetico utilizzo della feature selection, osservando solo i quattro modelli migliori presentati nella sezione antecedente, risulta difficile comprendere se in questo caso sia effettivamente utile: l'unico modo per rispondere è andare ad osservare tutti i dati ottenuti nel corso della trattazione e notare che in generale, anche nei casi in cui i risultati peggioravano in seguito all'applicazione della selezione delle caratteristiche, i cambiamenti non sono mai stati eccessivi, né in positivo né in negativo. In casi come questi, dove le informazioni sono molto limitate, anche nella pratica non si sente il bisogno di accorciare i tempi di elaborazione dei dati o di abbassare il peso computazionale, ma volendo generalizzare a casi molto più ampi e di conseguenza anche più affidabili si può affermare che l'utilizzo della feature selection possa essere consigliabile, visto che il suo contributo non ha provocato danni alle classificazioni.

In secondo luogo, è fondamentale parlare delle features che contribuiscono ad una migliore classificazione: se ci si fermasse anche in questo caso al modello risultato come il migliore, si dovrebbe affermare che solamente i parametri morfologici siano informazioni considerate utili all'ottenimento di un test diagnostico affidabile; lo scopo di questo lavoro però non è quello di selezionare il modello migliore, ma è quello appunto di osservare i differenti fenomeni nei vari scenari presentati. Alla luce di ciò, si può quindi affermare che i sintomi siano assolutamente da considerare per una migliore costruzione del modello perché, nonostante la struttura d'elezione non prenda in ingresso tali informazioni, è innegabile che nella maggior parte dei casi, l'utilizzo congiunto dei parametri morfologici e dei sintomi presenti alla diagnosi abbia permesso di migliorare i risultati ottenuti utilizzando solamente le informazioni geometriche.

Tutte queste conclusioni che sono state tratte dalle analisi e dalle ricerche finora fatte possono essere prese in considerazione e sviluppate ulteriormente: ad esempio si potrebbe provare a modificare e migliorare l'algoritmo di estrazione dei parametri morfologici in modo da renderlo più affidabile e capace di lavorare con qualsiasi tipo di immagine di risonanza magnetica senza avere limitazioni riguardo la qualità di esse; inoltre sarebbe altrettanto utile testare questi risultati con differenti tipi di reti neurali con differenti strutture: in questo caso è stata usata la struttura di tipo Feed Forward, molto semplice ma comunque efficace nel restituire informazioni indicative e valide sui risultati ottenuti, ma ciò non significa che sia necessariamente il modello migliore e più adeguato a questo tipo di analisi; ovviamente anche

il numero di classificatori è molto grande e si potrebbe provare a svolgere lo stesso processo con differenti strumenti che non sono stati inclusi in questo caso.

In conclusione, i risultati ottenuti devono essere considerati, utilizzati ed estesi per approfondire lo studio relativo alla malformazione di Chiari con l'obiettivo di ottenere dei mezzi utili ed affidabili per la diagnosi della patologia, la scelta della cura più adeguata, la scelta del corretto percorso chirurgico da seguire e, cosa più importante, per il benessere dei pazienti.

Bibliografia

Fric, R., Ringstad, G. & Eide, P. K. Chiari-malformasjon type 1 – diagnostikk og behandling. Tidsskrift for Den norske legeforening (2019) doi:10.4045/tidsskr.18.0455.

Bordes, S., Jenkins, S. & Tubbs, R. S. Defining, diagnosing, clarifying, and classifying the Chiari I malformations. Child's Nervous System 35, 1785–1792 (2019).

Ciaramitaro, P. et al. Diagnosis and treatment of Chiari malformation and syringomyelia in adults: international consensus document. Neurological Sciences 43, 1327–1342 (2022).

Carey, M., Fuell, W., Harkey, T. & Albert, G. W. Natural history of Chiari I malformation in children: a retrospective analysis. Child's Nervous System 37, 1185–1190 (2021)

Markunas, C. A. et al. Clinical, radiological, and genetic similarities between patients with Chiari Type I and Type 0 malformations. Journal of Neurosurgery: Pediatrics 9, 372–378 (2012).

Gad, K. A. & Yousem, D. M. Syringohydromyelia in Patients with Chiari I Malformation: A Retrospective Analysis. American Journal of Neuroradiology 38, 1833–1838 (2017).

Curone, M. et al. Chiari malformation type 1-related headache: the importance of a multidisciplinary study. Neurological Sciences 38, 91–93 (2017).

Chatrath, A. et al. Chiari I malformation in children-the natural history. Childs Nerv Syst 35, 1793–1799 (2019).

Botelho, R. V., Bittencourt, L. R. A., Rotta, J. M. & Tufik, S. Adult Chiari malformation and sleep apnoea. Neurosurgical Review 28, 169–176 (2005).

Dantas, F. L. R., Dantas, F., Caires, A. C. & Botelho, R. v. Natural History and Conservative Treatment Options in Chiari Malformation Type I in Adults: A Literature Update. Cureus (2020) doi:10.7759/cureus.12050.

Shah, A. H., Dhar, A., Elsanafiry, M. S. M. & Goel, A. Chiari malformation: Has the dilemma ended? J Craniovertebr Junction Spine 8, 297–304.

Shenoy, V. S. & Sampath, R. Syringomyelia. (2022).

Wang, J. et al. Acquired Chiari Malformation and Syringomyelia Secondary to Space- Occupying Lesions: A Systematic Review. World Neurosurgery 98, 800-808.e2 (2017).

Sharma, H., Treiber, J. & Bauer, D. Chiari 1 and Hydrocephalus – A Review. Neurol India 69, 362 (2021).

AISMAC. Associazione Italiana Siringomielia e Arnold Chiari.

Balasa, A. et al. Comparison of dural grafts and methods of graft fixation in Chiari malformation type I decompression surgery. Scientific Reports 11, (2021).

Erica Leila Ahngar Fabrik, La gestione neurochirurgica della Malformazione di Chiari di tipo I: il ruolo dell'Intelligenza Artificiale.

Falcicchio Antonella, Pasotti Andrea, Algoritmo per la segmentazione automatica di risonanze magnetiche cerebrali in pazienti affetti da malformazione di Arnold-Chiari.